UNIVERSITY OF CALIFORNIA,

IRVINE

P(Sleep) 2.0: Capturing the dynamics of nighttime sleep using probabilistic models and big data

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Psychology

by

Benjamin David Yetton

Dissertation Committee:

Associate Professor Sara C Mednick, Chair

Professor Michael Lee

Assistant Professor Bryce Mander

2019

DEDICATION

To

Mark and Lisa Yetton

&

Emily Shih

Table of Contents

List of Figures:

List of Tables

Acknowledgments

I would like to express my deepest appreciation for my doctoral advisor, Sara Mednick, whose passion for life and science encouraged me to continue to pursue, and eventually, thrive in the UCI/UCR Ph.D. program. I thank her for taking a chance on an unknown kiwi kid and invested countless hours of mentoring to develop my professional and academic skills. Her fun-loving and friendly nature allowed me to feel at ease communicating both crazy ideas and problematic issues. Her continual excitement for my work and a high degree of trust she placed in my non-traditional methods were extremely motivating. Without her guidance, this Ph.D. would not have been possible. Thanks for being a great mentor and friend.

I would like to thank Michael Lee for the many enlightening statistical conversations. His colloquial demeanor and clear explanations were a joy, and I always looked forward to a friendly Bayesian stats chat. I owe much of my knowledge and interest in data science to Dr. Lee. Cheers mate.

I would like to thank Bryce Mander for his guidance and feedback on this dissertation, and insightful previous work on sleep, spindles, and aging.

I would also like to thank other academic luminaries for their passionate teaching and guidance, namely Dr. Aaron Seitz, Dr. Dan Ozer, Dr. Jon Willits, and Dr. Christian Shelton. Also, thank you to the NSF, for their support of my work through an NSF GRFP fellowship.

This dissertation would not have been possible without the help of several CS undergrads, both at UCI and UCR. Thank you specifically to Juan Antonio Villagomez, Junbai Hou, Alyssa Darjuan, and Minjae Park. Finally, I would like to thank the members of the Mednick Lab for their continued support and insightful discussions. Thanks to Lauren Whitehurst, Negin Sattari, Lizzie McDevitt, Tenzin Tselha, Mohsen Naji, Mohamad Niknazar, Kate Simon, Pin-Chun Chen, Frida Corona, Jing Zhang, Kathrine Duggan, Lexus Hernandez, and Cody Raiza.

Curriculum Vitae

2019 **Ph.D., Cognitive Science**, University of California,

Irvine, USA

2016 **Masters of Arts, Cognitive Psychology**, University of California,

Riverside, USA

2010 **Bachelor of Engineering in Mechatronics** with First Class Honors, University of Canterbury, Canterbury, New Zealand

Abstract of the Dissertation

Benjamin David Yetton

Doctor of Philosophy in Cognitive Science

University of California, Irvine, 2019

Associate Professor Sara C Mednick, Chair

Sleep, a critical necessity for health, cognition, and well-being, consumes 25 years of the average human's life. Alterations in sleep's macro and microstate have been observed in multiple illnesses, including depression and Alzheimer's disease, and sleep may represent a biomarker of these difficult to diagnose pathologies. Despite this, adequate models of the dynamics of sleep able to capture the fine-grained minutia of sleep's many states and processes are lacking. The current work leverages over 1000 nights of healthy sleep polysomnography recordings from multiple open-source datasets. Data were automatically analyzed through a machine learning and data engineering pipeline to extract important features of the electroencephalogram, such as sleep spindles, slow oscillations (SO), and Rapid Eye Movement (REM) events. The counts of these sleep features and transition probabilities of sleep stages were modeled across the night in 30-second epoch intervals, using a flexible yet principled Bayesian cognitive modeling approach. The model, christened *P(Sleep) 2.0*, defines "typical" patterns of healthy sleep, and how they are affected by time, circadian processes and individual differences (e.g., age, sex). Transition probability dynamics were found to best be predicted by the previous 2 stages, time spent in stage, time across the night, age and sex, but not an age*sex interaction. Spindles, SO and REM events were also affected by the same features, but an age*sex interaction was additionally present. N2 spindles and both N2 and N3 slow oscillations were found to increase with time spent in a stage, whereas N3 spindles and REM decreased. Spindles and SO were more present in the evening compared to the morning; in contrast, REM increased over the night. REM counts and N3 spindles increased with

age, while N2 spindles and SO reduced. Females had more N2 spindles, and SO in both N2 and N3 than men. The current work replicates a body of previous literature based on small sample sizes, and the increased power in this study provides firm conclusions to arguments on the patterns of healthy sleep across the night, both at the macroscale (stages) and microscale (feature events). The proposed work has broader impacts on the detection of abnormal (and potentially pathologic) sleep patterns.

Key Sleep Terminology:

- *Sleep Stages:* The states that that brain transitions through across a period of sleep, clearly distinguishable from patterns in the Electroencephalogram.
- *Epoch:* A 30 second period of EEG, the traditional quantum of time used to classify sleep into stages
- *Bout:* A uninterrupted series of epochs of a specific sleep stage
- *Circadian:* The cyclic pattern of sleep phenomena across a 24hr period
- *Sleep efficiency:* The percent of time spent sleeping compared to the time in bed
- *Wake after sleep onset (WASO):* the minutes awake after sleep onset and before final waking
- *Total sleep time:* total minute asleep in one sleep session

Overview

Sleep makes up a 3rd of the average human's life and is impacted by many physical illnesses and almost all mental disorders. The quantification of sleep is, therefore, an important endeavor not just for scientific understanding, but for improving clinical outcomes. Recent years have seen the release of large open-source datasets, powerful machine learning models, more expressive modeling approaches (Bayesian networks, probabilistic models) and growing recognition for sleep as a diagnosis tool. This dissertation focuses on a newly developed model of sleep architecture and features termed *P(Sleep) 2.0* which leverages these recent advances. *Chapter 1* includes a general introduction on sleep stages and sleep features, and how they are affected by time, and demographics (such as age and sex). *Chapters 2 and 3* cover previous work by myself that were instrumental in the development of P(Sleep) 2.0 – A model of sleep architecture parameterized with Bayesian networks dubbed *P(Sleep) 1.0*, and the best-in-class Rapid Eye Movement detector developed by me in 2015. Finally, *Chapter 4* focuses on the methods and results of *P(Sleep) 2.0* and discusses its limitations and possible future directions. Chapters 2 and 3 summarize previous peer-reviewed works by myself (Yetton, McDevitt, Cellini, Shelton, & Mednick, 2018; Yetton et al., 2016), and finer methodological details can be found in their respective papers. Chapter 4 represents unpublished work.

Chapter One:

An introduction to sleep

Sleep is a dynamic, multi-dimensional process that reflects lifespan developmental changes in physical and mental health, as well as day-to-day state fluctuations. Alterations in stage patterns and durations (i.e., sleep architecture), are seen in almost all psychologic, neurologic and neurodegenerative disorders, including insomnia (Drake, Roehrs, & Roth, 2003), narcolepsy (Ferrillo, Donadio, De Carli, Garbarino, & Nobili, 2007), sleep apnea (McArdle & Douglas, 2001) and Alzheimer's (Mander et al., 2014), as well as depression (Kupfer, 1981) and schizophrenia (Wilson & Argyropoulos, 2012). However, not all deviations from prototypical sleep are indicators of pathology. Individual factors such as age (Ohayon, Carskadon, Guilleminault, & Vitiello, 2003), body mass index (BMI) (Taheri, Lin, Austin, Young, & Mignot, 2004), ethnicity and sex (Redline et al., 2004) contribute to sleep architecture, and differences are also reported after sleep deprivation (Daan, Beersma, & Borbély, 1984) or drug use (such as caffeine (C. Drake, Roehrs, Shambroom, & Roth, 2013), nicotine (Zhang, Samet, Caffo, & Punjabi, 2006), alcohol and marijuana (Cohen-Zion et al., 2009)). Quantifying the typical variability in sleep architecture and its relation to benign factors, as opposed to those which may be indicative of illness, is of clinical relevance.

Since the discovery of Rapid Eye Movement (REM) sleep in the 1950s (Aserinsky & Kleitman, 1953), sleep has been understood as a series of stage changes across the night. The American Academy of Sleep Medicine (AASM), currently classifies sleep into 4 stages: N1, N2, N3 and REM, each associated with a specific oscillatory profile in the Electroencephalogram (EEG) (Iber, Ancoli-Israel, Chesson, Quan, & others, 2007). These oscillation profiles represent the summation of synchronous firing from billions of neurons detected at each EEG electrode, and their pattern over time gives a gross measure of brain activity during each sleep stage. For example, REM sleep exhibits high power in the theta frequency band (4-8Hz), while delta activity (1-4Hz) is elevated in SWS. Additionally, time domain features such as

sleep spindles (a burst of thalamic brain activity in the 12-16Hz sigma range), K-Complexes (spontaneous sharp voltage spike followed by voltage depression), REMs (phasic rapid eye movements) and Slow Oscillations (SO: slow 0.4-1 Hz oscillations) are differentially present in each stage (Berry et al., 2012). To identify sleep stages, trained technicians visually inspect the time and frequency domain features of each 30-second window of multichannel EEG. While sleep's processes are continuous, these discrete 30-second "epochs" are a historical artifact of the 30-second page length of early paper-printed EEG systems. Over the 60+ years of sleep research, the proportions of time spent in each sleep stage has been well-defined. The proportion of N1 is relatively constant throughout the night, whereas N2, REM and N3 exhibit a time-dependence, with more N3 at the beginning of the night and a larger proportion of N2 and REM in the morning (Carskadon & Dement, 2005). Throughout the night, sleep is often interrupted with brief bouts of wake, known as Wake After Sleep Onset (WASO), which are more prominent in the morning (Yetton, McDevitt, Cellini, Shelton, & Mednick, 2018).

A common clinical and research practice is to assess sleep via gross summary variables such as the time in each sleep stage and sleep efficiency (time sleep vs total time in bed) along with the density (number per minute) of various sleep features. Most studies consider sleep as a unitary phenomenon and sum or average these sleep variables over the entire time in bed. However, sleep is dynamic and its stages progress in a semi-deterministic order across a period of sleep, generally beginning with Wake → Stage 1, then transitioning into a cyclic repetition of N2 → N3 → N2 → REM (Carskadon & Dement, 2005) (i.e. Ultradian cycles). Summarizing sleep architecture across the whole night can be useful for group-level differences, but it ignores the important temporal aspects of sleep stages and oscillations (Bizzotto, Zamuner, De Nicolao, Karlsson, & Gomeni, 2010). This issue is highlighted in Fig 1 (Yetton, Mcdevitt, Cellini, Shelton, & Mednick, 2018) in which two qualitatively different hypnograms are shown - the first exhibits very fragmented sleep, with shorter bout durations for each stage and a higher number of stage transitions compared to the second. Nonetheless, when using traditional stage

proportions as a measure, there is no quantitative difference. Thus, considering that sleep fragmentation is a marker of detrimental, but treatable, health disorders such as obstructive sleep apnea (El-Ad & Lavie, 2005), alternative measures and more complex models should be considered.

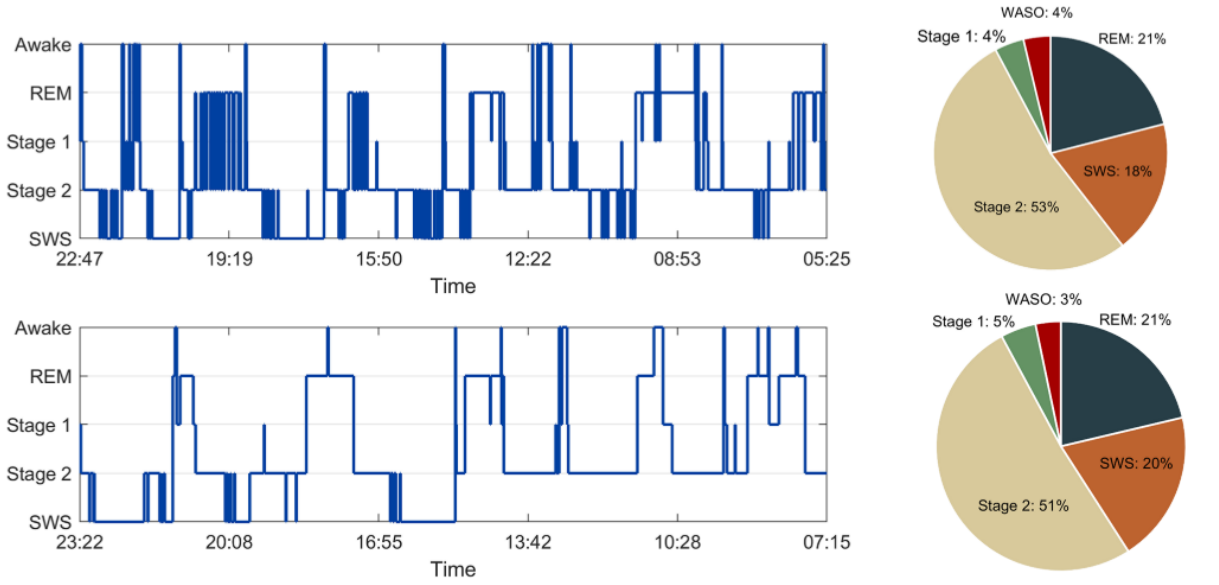

Figure 1: Hypnogram and corresponding state proportions of fragmented sleep (top) and normal sleep (bottom). Stage proportions are the minutes in each stage normalized by total sleep time.

It is now well established that distinct cognitive domains benefit from specific sleep stages and features. For example, our lab and others have shown that activity of sleep spindles (Mednick et al., 2013) and slow oscillations (SO: .5-1Hz, Rihms et al., 2014), as well spindle/SO coincidence (Niknazar et al., 2015) in a post-training sleep period correlate with the magnitude of declarative memory improvement (e.g., conscious, episodic memories), and spindles are also associated with improved attention (Cellini et al., 2015). On the other hand, minutes in REM sleep and REM theta activity (4-8Hz) correlate with improvement in non-declarative memories (e.g., unconscious, perceptual or sensorimotor skills) (Mednick, Nakayama, & Stickgold, 2003) and emotional salient memories (Groch, Wilhelm, Diekelmann, & Born, 2013; Hutchison & Rathore, 2015). Clinical psychopathologies can also be differentiated based on sleep features, with altered REMs patterns during sleep related to depression (Gillin et al, 1981; Lahmeyer et al, 1983; Mellman et al, 1997), narcolepsy (Vanková et al, 2001),

obsessive-compulsive disorder (Insel et al, 1982) and post-traumatic stress disorder (Ross et al, 1994; Mellman et al 1997, 2014) as well as changes in spindles and slow oscillations with depression, schizophrenia, and Alzheimer's disease (Mander et al., 2015; Manoach, Pan, Purcell, & Stickgold, 2016; Plante et al., 2013; Wamsley et al., 2012)) and sleep disorders (e.g. apnea (Carvalho, Gerhardt, Lemke, Schönwald, & de Santa Helena, 2012), narcolepsy (Bove, Culebras, Moore, & Westlake, 1994)). How the dynamical patterns of sleep stages contribute to cognition and healthy brain functioning is poorly understood. Models capturing these dynamics, therefore, have both scientific and clinical merit.

Like sleep stages, sleep features are also not independent of time. For example, Aeschbach & Borbély (1993) found a negative correlation between slow oscillation and spindle power, where the strength of this correlation reduced across the night. Within each period of NREM sleep, they noted opposite U-shaped time course for each feature, with high spindle density at the beginning and end of each cycle, and high slow-wave power in the middle. Merica and Blois (1991) found a slightly different pattern when considering band power, where, within an NREM episode, delta, theta, alpha, and sigma power began low, rose, peaked in the middle and then fell towards the end of the episode, but beta power displayed an inverted pattern. REM density also increases across the night (Peters, Ray, Fogel, Smith, & Smith, 2014), while SWA decreases with each consecutive NREM period (Achermann, Dijk, Brunner, & Borbély, 1993), and sigma power rising (Merica & Blois, 1991). Therefore, models to describe the dynamic process of sleep should not only account for the cycle of discrete sleep stages but also describe the variability of sleep features across time. This novel addition of sleep features to a model of sleep architecture would provide a starting point for the yet unknown relation between sleep feature dynamics and cognition and/or disorders.

Perhaps the historical lack of studies relating sleep features to each other and time is due to the large cost of obtaining this data. Currently, trained human experts are employed to detect sleep features, a process that is both time consuming and costly. With recent years and more powerful

algorithms, a glut of automatic sleep features detectors (Chambon et al., 2018; Lacourse, Delfrate, Beaudry, Peppard, & Warby, 2018; Massimini, 2004; Wamsley et al., 2012; Yetton et al., 2015) have been developed. This, coupled with an increase in freely available sleep data (Dean et al., 2016; O'Reilly, Gosselin, Carrier, & Nielsen, 2014; Yetton, Lacourse, Mednick, & Warby, n.d.) and stronger probabilistic modeling frameworks (Salvatier, Wiecki, & Fonnesbeck, 2016) create an opportune time for models of sleep that capture both stages and features.

Previous models of sleep architecture

Sleep has a long history of mathematical modeling. In 1982, Borbély proposed the dominate "two-process model" of sleep regulation (Borbély, 1982). It includes two separate mechanisms: the sleep/wake homeostat (Process "S") and circadian rhythm (Process "C"). Process S tracks sleep pressure (the drive to sleep), which rises (negative) exponentially during wake, and falls exponentially during sleep. Slow Wave Activity (SWA) levels in EEG (i.e. power between 0.5 and 4 Hz; often consider the same as Delta [1 - 4Hz]) track Process S during both wake and sleep (Achermann et al., 1993). Process C coordinates the light-dark cycle of day and night and cycles with a period of just over 24hrs. When Process S is high and C is low, the drive to sleep is high, likewise, when C is greater than S, a transition from sleep to wake occurs. Since its inception, the two-process model has been extended to account for REM/NREM "ultradian" transitions with the addition of a second periodic function (Achermann & Borbély, 1999; Achermann et al., 1993; Ferrillo et al., 2007). The two-process model is biologically plausible, parsimonious, and sufficient in accounting for global sleep phenomena, however, it is relatively simple (i.e. does not account for Stage 1, 2 and SWS separately) and cannot account for many of the individual differences seen in real hypnograms. Moreover, the sleep architecture pattern of real sleep is far from deterministic and often exhibits alternate patterns. Some of these deviations from typical patterns may be indicative of underlying mental or sleep disorders and deserve closer attention while others may be more benign. Quantifying just how much variance is expect for a healthy individual

helps to inform clinical tests on sleep and mental disorders. A model that captures the probabilistic, temporal pattern of all stages (N1-3, REM, and WASO), accounts for individual variability seen in normal sleep architecture (or abnormal sleep, if trained on abnormal data), and generalizes to new, unseen data, would have scientific and clinical relevance.

Probabilistic models of sleep dynamics have been proposed before. These models are generative (i.e. can simulate realistic a hypnogram), and explain both the average pattern of sleep as well as modeling the expected variance in these patterns. Kemp & Kamphuisen, (1986) simulated sleep hypnograms (sleep stage patterns) using a first-order Markov process model. Their parameterization assumes that the transition to a particular sleep stage *j*, from the current sleep stage *i* (where $j \neq i$), occurs with some probability that is either fixed across the whole night (p_{ij}), or dependent on the time (t) since sleep began ($p_{ij}(t)$, hand smoothed). The pioneering work of Kemp has been improved upon by allowing the transition rates from one stage to another to vary with time spent in a stage (christened *tau* - τ), i.e. $p_{ij}(\tau)$ (Kim, Lee, Robinson, & Jeong, 2009). There is some disagreement (Chu-Shore, Westover, & Bianchi, 2010) in the field over how to best parametrize p_{ij} with respect to τ (e.g. the form of duration distributions). Augments have been made for piecewise exponential (Kim et al., 2009) or multi-exponential (Bianchi, Cash, Mietus, Peng, & Thomas, 2010) for all stages, for exponential or stretched exponential for Stage 1, 2 and REM and power laws for SWS and Wake among others (Kishi et al., 2011; Kishi, Struzik, Natelson, Togo, & Yamamoto, 2008; Kishi, Yamaguchi, Togo, & Yamamoto, 2018). More complex models have assumed that a set of unobserved latent states generate the duration distribution for each stage (Bianchi et al., 2012). Along with previous stage information, transition probabilities have been coded as functions of other variables such as the time since bedtime (Bizzotto et al., 2010; Karlsson et al., 2000), or the wall clock time (to account for circadian effects)(Yetton et al., 2018).

Dynamics of sleep

The pattern of sleep stages over time is semi-deterministic. That is, the likelihood of the current stage and its duration is a probabilistic function of the stages that came before it, their durations, as well as factors such as time of day. It is clear from this past modeling work and existing literature (De Oliveira Alvares et al., 2013) that some sleep stage transitions are more or less likely (e.g. N3->REM is unlikely, while N3->N2 is likely), therefore the identity of the current sleep stage influences the next stage. But how far back in the history of stages should be considered when modeling sleep architecture? My previous work determined that adding stage information from 3 stages back did not improve the prediction of the next stage. Likewise, work by Kishi et al (Kishi et al., 2018), found that accounting for the previous 2 stages was sufficient to replicate the well-known 90-minute ultradian rhythm. Therefore, the current *P(Sleep) 2.0* model includes only the previous 2 stages.

The two-process model, and experimental observation (Carskadon & Dement, 2005), have shown variation in sleep architecture based on circadian timing (*Time Of Day*) and magnitude of sleep pressure (*Time Slept*). Like the two-process model, my previous work also found an effect of these variables. What is surprising, however, is that *Time Slept* and T*ime of Day* were not needed to predict the next stage and its duration *when at least the two previous stages were considered as predictors*. I suggested this is likely because specific patterns of stages (and their durations) occur at different times over the hours of sleep, and once a specific pattern has been initiated, then *Time of Day* provides no extra information to predict the pattern end (i.e. the previous two stages already capture information related to both time slept and time of day). Further, it suggests that a model parameterized by transition probabilities has enough capacity to capture longer these long term temporal patterns. However, this finding may also be explained by the low resolution of my previous model. Further, without circadian desynchronization protocols, these variables are highly collinear (both increase across the night, and differ by the time spent in wake only), and the inclusion of one is likely sufficient.

Consequently, P(Sleep) 2.0 still includes *Time of day* (i.e. clock-time) due to its strong theoretical bases and this model's finer grain continuous framework, where stages, their duration, *and sleep features* are predicted at an epoch by epoch level.

Sleep and Individual Differences

Sleep architecture is variable across the population, and some of this variability may be explainable by demographic measures (such as age, sex, and BMI). Likewise, demographics also explain variance in sleep features.

Age effects on sleep patterns

Age is a strong predictor of sleep patterns. Older adults have difficulty entering and maintaining sleep, with increased fragmentation of stages (shorter bouts) and a greater number of transitions, especially to wake. Conte et al (2014) describe the more variable pattern of sleep as increased "functional uncertainty", and point to a deterioration of the central nervous system in consolidating and maintaining coordinated physiological processes. They also raise the intriguing hypothesis that the reduction in social and cognitive demands directly leads to a reduction in sleep continuity, pointing to decreases in fragmentation after cognitive training in the elderly (Conte, Carobbi, Errico, & Ficca, 2012). Focusing on sleep stages specifically, older adults spend more time in Stage 1 and 2, and trade this off with less time in SWS; the effects of age on REM are less clear (Mander, Winer, & Walker, 2017; Ohayon et al., 2003; Redline et al., 2004). My work found substantial effects of age on sleep architecture dynamics, influencing both transition probabilities and bout durations. Specifically, I found non-linear age effects whereby SWS minutes, proportions, durations, and transition probabilities reduced sharply between youth and mid-age but flattened off in later life. For WASO, the opposite was true, this stage was more constant up until mid-age and then increased abruptly.

Age effects on sleep features

Age has drastic effects on spindles and slow oscillations. Recent work by myself and colleagues investigating spindles across age and sex found that spindle density, duration, and amplitude all decrease with age and that females had greater density and amplitude compared to males (Yetton et al., n.d.). Slow oscillations also decrease with age, with reduced power, density and amplitude (Carrier et al., 2011; Dubé et al., 2015). These findings replicate a large body of previous literature (Mander et al., 2017; Peters et al., 2014; Purcell et al., 2016; Sattari et al., n.d.; Ujma et al., 2015). Topological changes are also present with age (Martin et al., 2013). The drastic age changes, especially in metrics important for feature detection (such as duration and amplitude), may lead to spindle and slow oscillation detectors reporting lower spindle counts. These lower counts are not because fewer spindles exist per se, but simply because the poor signal to noise ratio in older adults limits detectability (although both may be possible). I chose to restrict spindle/SO analysis to 40 years and below (a cutoff also chosen by (Purcell et al., 2016)) such that poor feature detector performance would not bias results. For REM events/density, changes with age and sex are less clear. Only a handful of studies investigated REM density changes with age, with some reporting lower REM density with age (Darchia, Campbell, & Feinberg, 2003), while others found no differences (Ficca et al., 1999; Schwarz et al., 2017). Very few studies considered sex effects on REM density, however, 2 studies found an interaction of sex and age, where REM density increased over age for males and decreased for females (Reynolds III et al., 1985; Roch, Reynolds III, Kupfer, & Berman, 1988). Given the limited consensus on rem density vs age changes, I chose not to age limit REM density analysis in *P(Sleep) 2.0*.

Sex effects on sleep

Sex also moderates sleep patterns. Men tend to have lighter (more Stage 1 and 2) sleep than women, however, sex effects are most apparent though its interaction with age. In a smaller sample, Ehlers & Kupfer (1997), demonstrated greater Stage 2 and decreases in both REM and SWS in between

men aged 20-30, and 30-40 years, while no effects were seen in women. In a larger dataset (partially included in Yetton et al, 2018), Redline et al, (2004) showed older females had, on average, 106% more SWS than age-matched males. The sex hormone testosterone plays an import part in healthy, consolidated sleep (Mander et al., 2017) where reduced levels of testosterone lead to reductions in SWS (Latta et al., 2005). Older males exhibit reduced levels of testosterone, and this may contribute to shorter durations, increased fragmentation of SWS and increased transitions to wake for older males (Yetton, McDevitt, Cellini, Shelton, & Mednick, 2018). Furthermore, in some females, sleep disruption increases after menopause (Baker, Willoughby, Sassoon, Colrain, & de Zambotti, 2015). Postmenopausal women also exhibit increased SWS (Young, Rabago, Zgierska, Austin, & Finn, 2003), which may counteract the main effects of age or even help promote an uptick in SWS in older females. My own work also found interactions between sex and age, in both traditional sleep measures (time spend in a stage) and dynamic measure of SWS and WASO. Males had greater deficits in SWS as they age, exhibited by less total minutes in this stage, and reduced transition probabilities and durations. Given theoretical underpinning and the finding from my previous work, *P(Sleep) 2.0* considered sex and sex*age interaction terms.

BMI effects on sleep

Rao *et al.* (2009) reported increased BMI associated with a decreased proportion of SWS in an older population (MrOS Study) after controlling for age, sex, clinic location, race, sleep efficiency, sleep-disordered breathing, and other health variables. Additionally, Redline *et al.* (2004) found increased lighter sleep stages and reduced SWS associated with low BMI. My own investigation did not find any BMI effects (Yetton, Mcdevitt, Cellini, Shelton, & Mednick, 2018), and therefore I did not include them in P (Sleep) 2.0 (although BMI limits were in the subject inclusion criteria).

Chapter Two:

P(sleep) 1.0 – A Bayesian Network Model of sleep architecture

Using big data and modeling techniques borrowed from artificial intelligence (Discrete Bayesian Networks) I captured the expected progression of sleep stages across a night. Bayesian network analysis began with model selection algorithms to determine which variables must be considered when modeling sleep architecture dynamics, and which can be safely excluded to reduce model complexity. Next, the best fitting model was used to determine how sleep changes across the night. The dataset used was large (~3200 subjects) and consists of full night recordings from individual subjects across the full range of demographics from a healthy population.

The pattern of sleep stages over time is semi-deterministic. That is, the likelihood of the current stage and its duration is a probabilistic function of the stages that came before it, their durations, as well as factors such as time of day. I sought to determine the temporal structure of sleep architecture by testing the influence of previous stages (both stage *identity* and duration) on the current stage. The data I was trying to model was therefore a sequence of {stage identity, duration} pairs (e.g. {Stage 1 for 5 minutes} -> {Stage 2 for 3 minutes} -> {SWS for 20 minutes}), where each data-point is about of sleep. However, my use of discrete Bayesian networks required discretization of duration from continuous minutes into 4 duration bins (e.g. if bout duration shorter than 1.5 mins, assign 1, if between 1.5 and 5 mins assign 2, etc). I predicted each bout (both identity and duration) given more or less previous stage information. The Bayesian network was therefore parameterized by 2 types of probability tables: 1) transition probability tables, that gave the probability of the next stage identity given previous stage and duration information, and 2) duration distribution tables, which gave a low-resolution probabilistic representation of the duration of the next stage. Both these parameterizations have been used

previously to quantify sleep dynamics (Bianchi et al., 2012; Chu-Shore et al., 2010; Kemp & Kamphuisen, 1986; Kishi et al., 2008, 2011; C.-C. Lo et al., 2004; C. Lo et al., 2002; Schlemmer, Parlitz, Luther, Wessel, & Penzel, 2015; Wei et al., 2017).

We know that the current stage (stage at *t-1*) influences the next stage (stage at *t*), however, does the identity of both the previous stage (stage at *t-2*) and current stage influence the next stage? How far back in time to we need to consider? To answer this question, I used the K2 structural search algorithm (Cooper & Herskovits, 1992) to find the best-fitting Bayesian network over sleep architecture variables. This algorithm starts with a network with all variables independent (representing no relation between variables) and adds connections between variables only when the Bayesian Information Criterion (BIC) does not increase. Three models were tested with varying amounts of previous stage information included: one back (*t-1*), two back (*t-2*) and 3 back (*t-3*) (Figure 2).

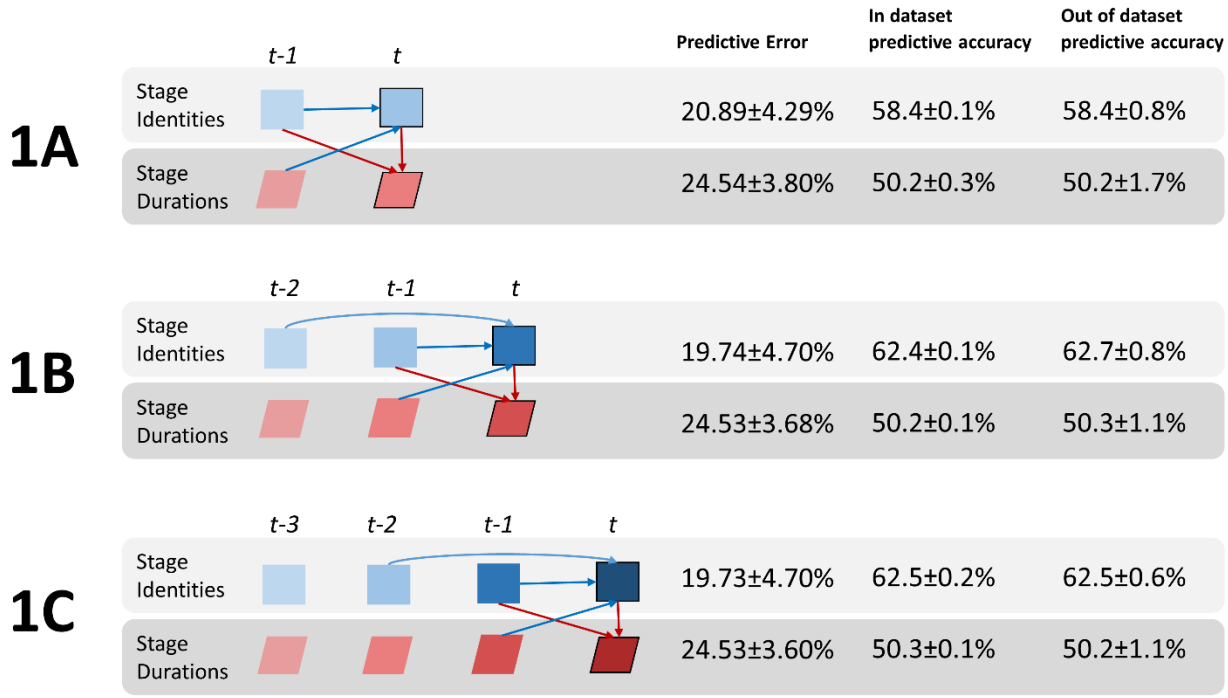

Figure 2: Best fitting models to predict the current stage and duration from previous sleep architecture variables. "In dataset" and "out of dataset" prediction accuracy and prediction error for current stage (top) and current stage duration (bottom) are shown. ***1A****) 1 back model (including t-1 variables),* ***1B****) 2 back model (including t-2 variables),* ***1C****) 3 back (including t-3 variables). When considering previous sleep architecture only, Model 1B gave the best fit*

and states that the identity of the current stage is dependent on the identity of the previous 2 stages and the duration of the last stage (blue arrows). The duration of the current stage is dependent on the identity of the current stage (at t) and the previous one (at t-1) (red arrows).

Comparing Figure 2**,** Model 1B and 1C, I saw the identity of the current stage (at *t*), is optimally predicted by the two stages before it and the previous stage duration (at *t-1*). The duration of the current stage (at *t*) was a probabilistic function of the identity of the current stage (at *t)* and the previous stage (*t-1)*. The best model (1B) predicts the next stage at 62.5% accuracy and the duration of the next stage at 50.3% accuracy (chance at 20% and 25% respectively). Out of dataset accuracy was no different than in dataset accuracy, suggesting high model generalizability. Adding stage information from 3 stages back (i.e. t-3, Model 1C) did not aid predictive power (no connections from t-3 to t). Hence, I concluded the probability of the current stage may be modeled as a 2^{nd} order Markov Process and its duration a first-order Markov Process.

The influence of time of day

Circadian processes and the percentage of time already slept influence sleep architecture. To the Bayesian network models, I added the *Time of Day* and *Time Slept* variables (discretized) to test their influence on the current stage and its duration. Three models were run: without any previous stage information, with 1 back stage information (current stage, t-1), and with 2 back stage information (current, t-1, and previous stage, t-2) (Figure 3, 2A, 2B and 2C).

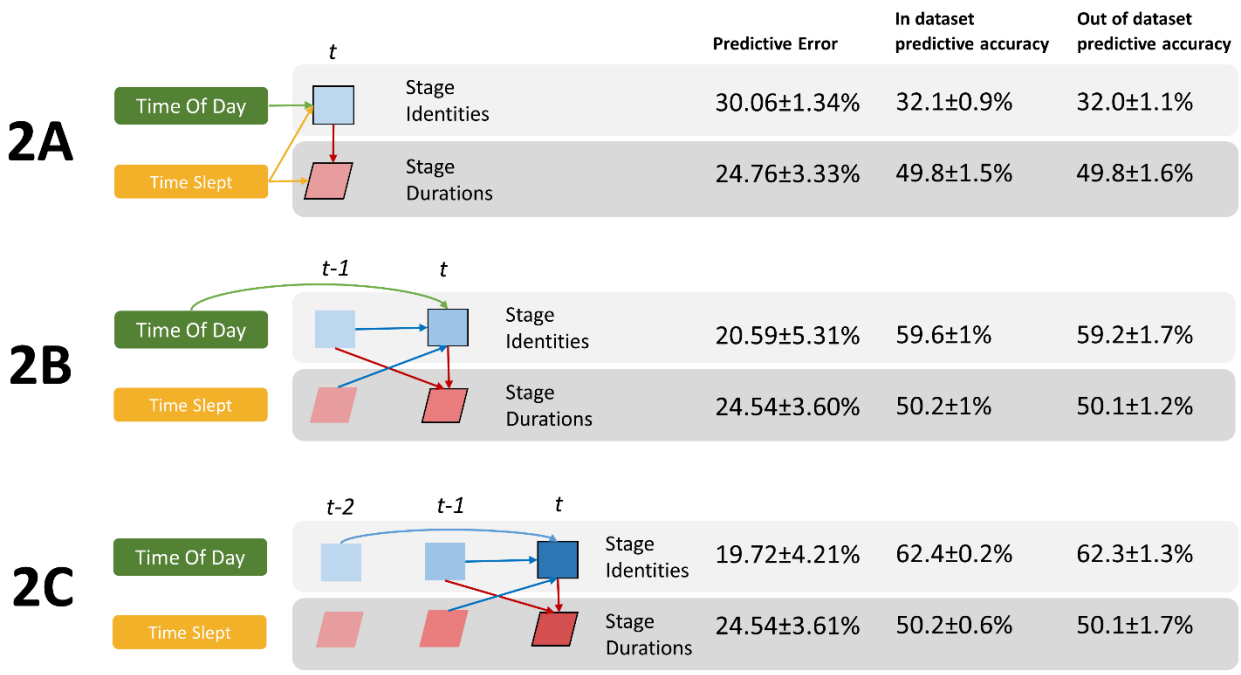

Figure 3: Effects of Time of Day and Total Sleep Time. ***2A****: When previous stages not included,* ***2B****: 1 back included,* ***2C****: 2 back included. Beside each model is the "in dataset" and "out of dataset" prediction accuracy and prediction error for the current stage (top) and current stage duration (bottom). Time of Day influences 0^{th} order transition probabilities (2A) and 1^{st} order transition probabilities (2B). Total Sleep Time influences both 0^{th} order transition probabilities and duration distributions when no previous stage information is available (2A).*

I found that *Time of Day* affected the likelihood of the current stage when no previous stage information is available (i.e. 0^{th} order transition probabilities, 2A) and the probability of the current stage given the last (i.e. 1^{st} order transition probabilities, 2B). *Time of Day* did not aid the prediction of the current stage when at least the last two stages were known (2C). *Time Slept* influenced both the current stage and its duration, but only when no previous stage information was available (2A). The finding of model 2A and 2B are not surprising; the two-process model, and experimental observation (Carskadon & Dement, 2005), have shown variation in sleep architecture based on circadian timing (*Time Of Day*) and magnitude of sleep pressure (*Time Slept*). In addition to the two-process model, this model suggests that stage duration changes are influenced by the duration of time already spent sleeping, rather than circadian effects.

What was surprising, however, is that *Time Slept* and T*ime of Day* were not needed to predict the next stage and its duration *when at least 2 stages back were considered (2C)*. This is likely because specific patterns of stages (and their durations) occur at different times over the hours of sleep, and once a specific pattern has been initiated, then *Time of Day* provides no extra information to predict the pattern end. The 2C model result may be due to the low resolution caused by the discretized view of sleep. However, a recent Markov chain model (Kishi et al., 2018) corroborates my finding and demonstrates that a 2 back model without a time of day effect is sufficient to capture sleep architectures ultradian patterns (NREM/REM alternations).

Individual difference measures on traditional sleep variables

Using the same 3202 subjects as in Bayesian network analysis above, I fit multi-level regression models ("dataset" as level 2), to predict minutes in each stage in a stepwise procedure from age, sex, and higher-order terms while controlling for total sleep time (TST), and sleep onset time. While these analyses fail to capture dynamic trends, they remain useful as global summary statistics and for comparison with previous literature (Ohayon et al., 2003), and highlight the gross effects of age and sex. Each relationship is shown graphically in Figure 4 (where TST and sleep onset are fixed at their mean). For all relationships, a random slopes model better accounted for the data compared to a pooled or random intercept model.

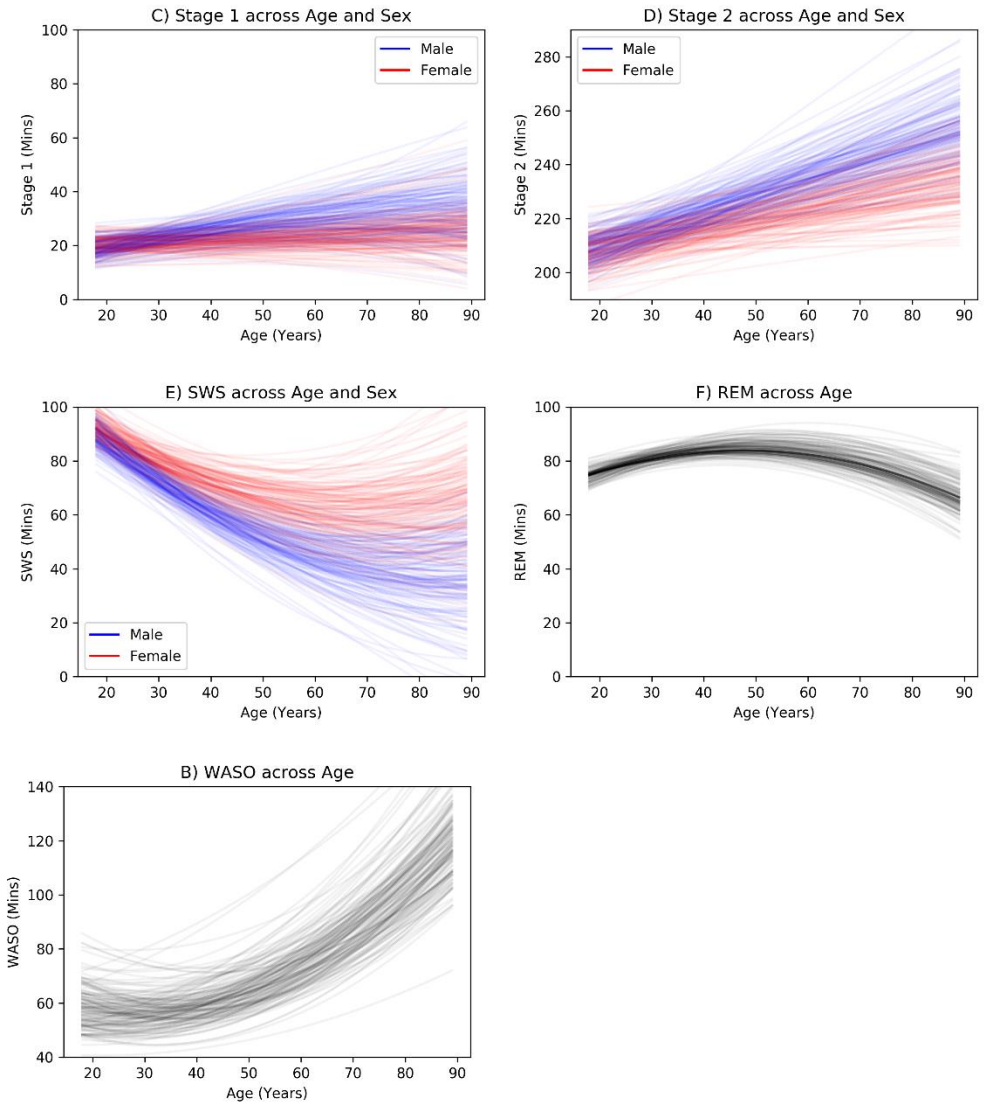

Figure 4: Minutes in each stage across age and sex. To show the uncertainly in predictions, regression parameters are randomly sampled 100 times from each model's joint parameter distribution and each is used to plot a regression line. A plot with grey lines represents no sex effect. REM: rapid eye movement sleep; SWS: slow-wave sleep; WASO: wake after sleep onset.

For stage 1, the model with age, sex, age*sex, age^2 and the age^2*sex interaction best explained the data (P(Correct Model | Data, Models)=0.87; P(next best model, without age^2*sex)=0.08). For females, minutes in stage 1 increased only very slightly throughout their lifes (B_{age}= 0.06, CI=[-0.36, 0.50], B_{age2} = 0.000 [-0.004,0.004]). Males, on the other hand, increased by over half a minute for each year of life (B_{age} = 0.37, 95%CI = [-0.13,0.89]), however, this increase slowed somewhat with age (-0.002, 95%CI = [-0.008,0.003]).

Stage 2 was best predicted from age, sex and the age*sex interaction (P(Correct Model | Data, Models)=0.85; P(next best model, inc. age^2)=0.07). Minutes in stage 2 began around 210 minutes for both sexes, and increased 39 seconds per year for females (B_{age}=0.39, 95%CI = [0.09,0.76]), and faster, at over 1 minute per year for males ($B_{age*sex}$ = 0.26, [-0.05, 0.57]).

For SWS, age, age^2, sex and the sex*age interaction were selected in the best model (P(Correct Model | Data, Models)=0.86; P(next best model, inc. sex*age^2)=0.14). Both sexes spent less time in SWS as they age (B_{age}= -1.67, 95%CI = [-2.36, -0.91]), but this effect reduced and began to flatten off at ~70 years (B_{age} = 0.01, [0.00,0.02]). Males had less SWS than females, and had a sharper reduction of SWS with age than females ($B_{age*sex}$= -0.34, 95%CI = [-0.66, -0.07]).

The best model for REM included age and age^2 as predictors. This relationship was interesting, in that REM minutes began low, and increased through to mid-age, and then began to decrease again (B_{age} = 1.10, 95%CI = [0.76,1.53], B_{age2} = -0.01 [-0.02, -0.01]).

For WASO, a model with age and age^2 variables, but not sex was selected (P(Correct Model | Data, Models)=0.76; P(next best model, without age^2)=0.22). Here I found that WASO increases non-linearly across age, beginning at around 60 minutes at 18 years, remaining flat until 45 years, and then increasing sharply throughout middle and older age (B_{age} = -0.94, 95%CI = [-1.84, -0.19]; B_{age2} = 0.02, 95%CI = [0.01,0.02]).

Taken together, these results suggest age affects more sleep stages than sex. Further, as men age, they spend more time in Stage 1 and 2, and this is offset with less time in SWS. For females, the

same pattern exists, but changes across age are less pronounced. For both sexes, WASO minutes remain constant in early life, but then begin increasing at around 45 years of age. REM minutes showed no changes across sex but a curvilinear relationship was observed where mid-age adults have the most REM. The rise in SWS in older age, which was more pronounced in women is intriguing. This is potentially an artifact of mortality, where the only surviving 80-90-year-olds are those with higher SWS.

Individual differences in the dynamics of sleep architecture in P(Sleep) 1.0

To investigate the effects of individual differences on the dynamics of sleep, I added individual variables of *Sex, Age* and *BMI* to the Bayesian network models from above (Figure 5). I found BMI had no influence on stage durations or transition probabilities (when *Time of Day* and *Time Slept*, *Sex* and *Age* were accounted for). *Sex* modulated the probability of the current stage's identity (i.e. different 0^{th} order transition probabilities for each sex group). *Age,* modeled categorically as younger (18-42 years), middle-aged (43-66 years), and older (67-90 years) groups, had substantial effects on sleep architecture dynamics, influencing 0^{th} order transition properties (Model 3A), and influenced stage durations even when all previous stages that were predictive were included (3A, 3B, 3C).

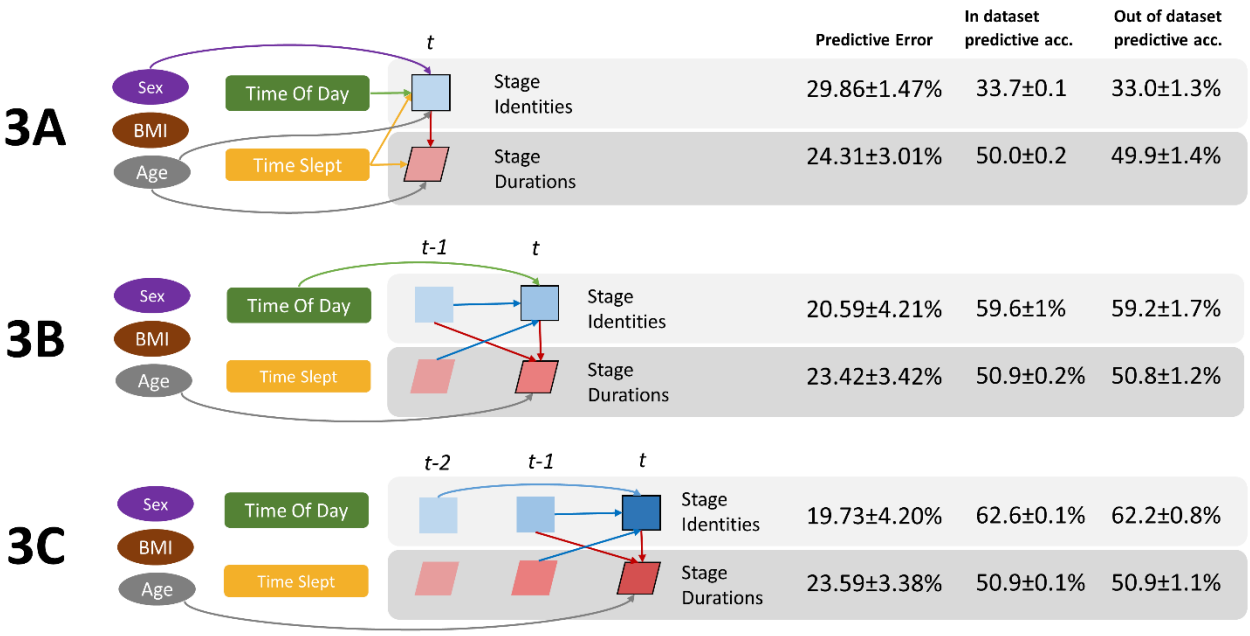

Figure 5: Full model including individual factors: ***3A:*** *No previous stage information,* ***3B:*** *1 back stage information,* ***3C:*** *2 back stage information. Beside each model is the "in dataset" and "out of dataset" prediction accuracy and prediction error for the current stage (top) and current stage duration (bottom). BMI: Body Mass Index.*

Consideration from P(Sleep) 1.0 for P(Sleep) 2.0

Contrary to previous literature, the study found no relationship between BMI and sleep architecture. Potential reasons for this discrepancy may be that prior studies had less stringent BMI exclusion criteria, looked at stage proportions rather than duration distributions and transition probabilities, and summed SWS over the whole night. These differences may account for the alternative BMI findings. BMI information was missing from some studies, and the lower number of data points may have contributed to the lack of a BMI finding. However, after removing data without BMI information and running all models without BMI, the best fitting relationship among variables was unchanged for all, suggesting that a lack of statistical power was not an issue. Therefore, *P(Sleep) 2.0* did not include BMI.

Similar to previous work (Ehlers & Kupfer, 1989; Mander et al., 2017; Ohayon et al., 2003), in Yetton et al 2018 I found large effects of age, particularly on SWS and WASO. My Bayesian network and regression analysis both found non-linear age effects for these stages whereby SWS minutes, proportions, durations, and transition probabilities reduced sharply between youth and mid-age but

flattened off in later life. For WASO, the opposite was true, this stage was more constant up until mid-age and then increased abruptly (for both minutes + durations). Given these age effects, age must be was considered as a variable in *P(Sleep) 2.0*. The exact form of the age to sleep architecture relationship was not apparent from this work, but a model that includes the quadratic effects of age should certainly be investigated, given the strong quadratic effects seen in the analysis of traditional measures.

It is interesting that sex was only found to influence transition probabilities when no previous sleep stage information was included. The lack of a sex influence in higher-order models may be due to discretization of sleep or small effects of sex, where extra parameters that account for sex did not increase model fit enough to counteract the increase in model complexity (The BIC measure used trades off between model fit and parsimony). I, therefore, included sex in P(Sleep) 2.0.

The expected interactions between sex and age were also apparent, particularly in both static and dynamic measure of SWS, and dynamic measures of WASO. Males had greater deficits in SWS as they age, exhibited by less total minutes in this stage, and reduced proportions, transition probabilities, and durations. Given theoretical underpinning and the finding from *P(Sleep) 1.0*, *P(Sleep) 2.0* considered sex and sex*age interaction terms.

Chapter Three:

Sleep features and their detection via automatic algorithms

Critical for understanding and modeling sleep is the development of automated feature detection tools. Automated methods do not fatigue, are cost-efficient, reproducible, and are readily deployable. While expert human judgment is still considered the clinical gold standard, a number of automated methods to detect spindles (Lacourse et al., 2018; Wamsley et al., 2012), REMs (Hatzilabrou et al., 1994; Yetton et al., 2015), K-complexes (Lajnef et al., 2015), and slow oscillations (Massimini, 2004) have been proposed.

It should be noted that the sleep features detected by humans and machines differ. Compared to humans, automated detectors have a high false-positive rate and tend to overestimate the number of occurrences of a sleep feature (Lacourse et al., 2018). This is especially true with spindles, which are often masked by other signal phenomena (such co-occurrence with slow waves). However, Lacourse et al note that these detectors often find "spindles" in wake, REM and stage 1, which is problematic considering the current definition of spindles restricts them to stage 2 and SWS only. Further, different automated detectors find different spindles (Warby et al., 2014), whereas the scoring across humans is more consistent. Another method to emulate human scoring is to optimize scoring algorithm parameters such that the argument between humans and machines is maximized. This chapter describes one such automated method (Yetton et al., 2015) which employs machine learning to detect Rapid Eye Movement Events within REM sleep.

Given that the average human spends approximately 5-6% of her life in REM sleep, it is surprising how little investigation has been undertaken into this psychophysiological event. This is due,

in part, to the difficulty, and time-consuming task of manually counting REMs (see Figure 6 for examples). Development of a computational algorithm to automatically detect these ocular events would provide a solution to this problem, and open a pathway to aid in the discovery of the nature and function of REMs. Here, I provide an automated approach to REM detection (examples shown in Figure 6) utilizing a learning algorithm designed to detect REMs from the Left Ocular Canthi (LOC) and Right Ocular Canthi (ROC) channels of PSG data. I compared the novel algorithm to several prior published algorithms (Agarwal, Takeuchi, Laroche, & Gotman, 2005; Doman et al., 1995; Hatzilabrou et al., 1994; Ktonas & Smith, 1978; McPartland, Kupfer, & Foster, 1973; Minard & Krausman, 1971) and to expert and non-expert human scorers. The best performing single algorithm integrates a novel set of features and the powerful 'AdaBoost' classification algorithm to detect the presence of REMs.

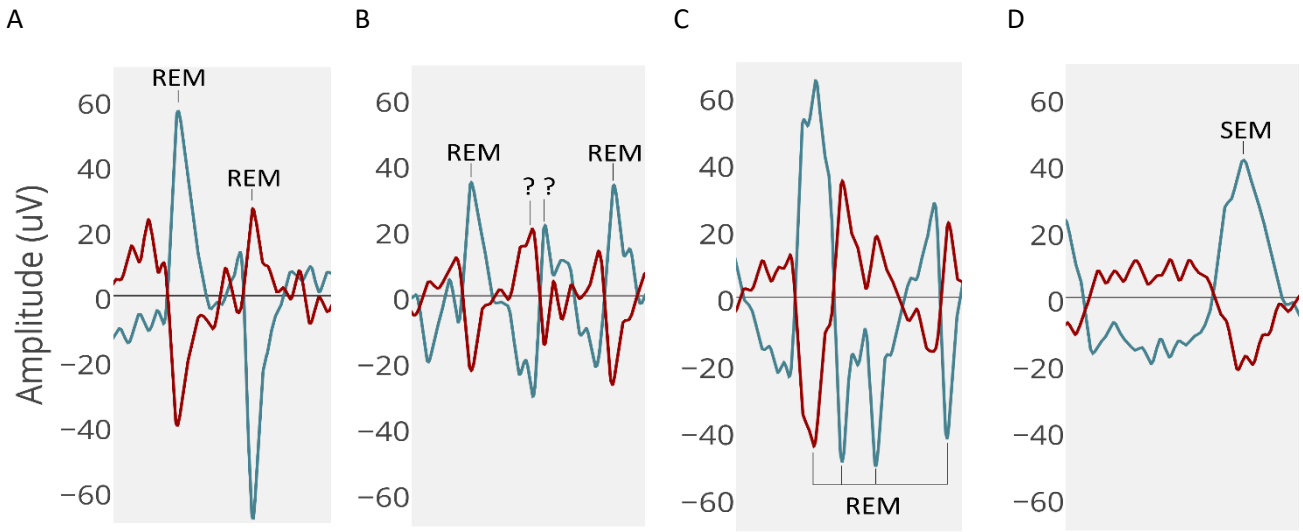

Figure 6: Examples of REM waveforms in LOC and ROC channels. Y-axis represents 3 seconds (3 windows). LOC in blue, ROC in red: ***A.*** *An example of two 'ideal REMs' easily detected by simple thresholding* ***B.*** *Example of REM-like movements (?) to be ignored* ***C.*** *Multiple REMs in close proximity with different amplitudes in channels, requiring a combination of features to detect* ***D.*** *Slow Eye Movements (SEM) to be ignored.*

Methods

Data Set

The data set consisted of 5 (3 Female) subjects' polysomnography data (Astro-Med Grass Heritage model 15 amplifiers [Natus Neurology Incorporated, West Warwick, RI, USA]) taken from the control condition of a previous nap study (Mednick et al., 2013). Subjects gave informed consent and the study protocol was reviewed by the University of California, San Diego Institutional Review Board. Participants were healthy (BMI = 23 ± 2.4 kg/m^2), non-smokers aged 18 to 35 (22 ± 3 years) with no personal history of neurological, psychological, or other chronic illness and were normal sleepers, habitually obtaining approximately 8 hours of sleep each night. Informed consent and original study was.

American Association of Sleep Medicine (2007) guidelines recommend two channels of EOG for REM detection, "recording from an electrode placed 1 cm above or below the outer canthus of the eye" (LOC and ROC). Although the placement of these channels does not allow for the ability to detect differences between vertical and horizontal eye movements (Varri, 1996), they continue to remain the gold standard eye channels used in sleep labs. The goal of my algorithm is to generalize to the majority of sleep labs, hence LOC and ROC are used.

REMs Scoring

Expert sleep scorers identified 110 minutes of REM stage sleep and this was used to train and test the algorithm. REM peaks in each subject's PSG data were independently identified by an *expert group* (3 expert sleep scorers, each with 2 or more years of experience) and a *non-expert group*, (4 non-expert sleep scorers, familiar with PSG, and having undergone basic in lab training on identifying REMs) by marking REM movement peaks. Raters adhered to the AASM (2007) REM definition of 'conjugate, irregular, sharply peaked eye movements with an initial deflection usually lasting less than 500 msec'. Horizontal lines at ± 37.5uV (as suggested by Werth, Dijk, Achermann, & Borbély, 1996) were used as

visual aids to alert raters to the possible presence of REM, although the EOG signal did not have to cross these lines to be considered as a REM. Slow Eye Movements (SEM) with sinusoidal peaks and longer deflection times (> 500ms) (AASM, 2007) were not considered by raters and are to be ignored by my algorithm.

Algorithm Development

Overview

I employed two approaches: feature thresholding and machine learning. I obtained respectable results using a hand-tuned threshold-based approach with an intersection of Amplitude, Slope, Cross-Correlation, and Discrete Wavelet Transform (DWT) (e.g. if *Amplitude > w* AND *slope > x* AND *Cross-Correlation > y* AND *DWT > z* then label as a REM). By adding the features with high recall first and only adding precision in the later steps, I began with a high number of true and false positives and iteratively reduced false positives while controlling for false negatives.

A large set of features can be predictive, and this simple thresholding approach becomes intractable when REMs are best predicted from multiple feature interaction terms. In the second approach, I used an adaptive boosting (AdaBoost) classification algorithm (Freund & Schapire, 1999), which is able to automatically tune the thresholds and combinations of multiple features by learning the statistical regularities that predict REM from a training data set. It combines many simpler models, each of which focuses on different examples of REM. This algorithm has been previously used on EEG signals to successfully classify epilepsy-related EEG signals (Niknazar et al., 2013) and schizophrenia (Boostani, Sadatnezhad, & Sabeti, 2009).

Our AdaBoost detection algorithm followed a 4-step process (Figure 7). First, data was filtered, discretized (by splitting into consecutive windows) and a gold standard was created from human scorers (correct classification of each window). Next, features were extracted from each window. Third, a

classifier algorithm trains a classifier to learn to distinguish between windows containing no REM, a single REM or two REMs in an iterative manner. Finally, the testing set was run through the now trained classifier, and classification performance was measured by comparing classifier output to the expert human gold standard.

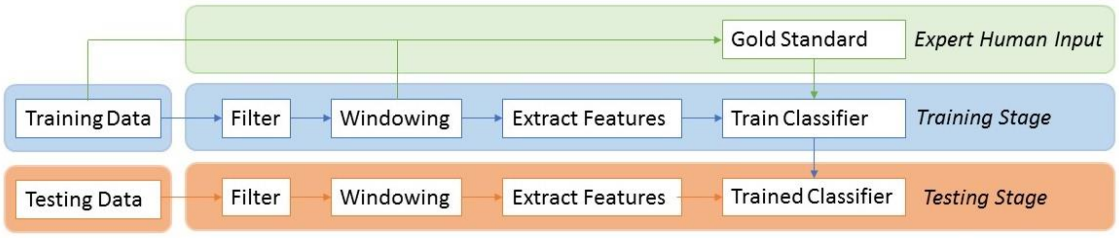

Figure 7: Algorithm Overview.

Filtering and Windowing

The spectral power across all REMs peaks between 0.3Hz and 5Hz. Frequency components higher or lower were considered as noise and removed with a zero-phase digital bandpass filter [0.3Hz - 5Hz, 40DB attenuation]. Other filter cutoff frequencies ranging from 0Hz to 15Hz were considered but did not improve performance. Other than initial bandpass filtering, EOG signal artifacts were not removed from the dataset and did not affect results. The filtered LOC and ROC data across subjects were divided into 8022 consecutive 1-second windows, each of which undergoes feature extraction. This window size was chosen to capture the average time for a complete REM movement. Results of other window sizes (0.5, 0.7, 1.2, 1.4, 1.6 seconds among others) were investigated, but performance did not increase.

Gold Standard

Raters were instructed to mark REM peaks by hand. A REM peak location 'gold standard' was created by combining the REM peaks scored by each rater. To avoid counting the same REM as identified by different raters, MATLAB's hierarchical clustering algorithm (The MathWorks Inc., 2015) was used. Effectively, REM peak marks closer than 120 ms across raters were merged into a single REM

with the resulting peak at the maximum absolute value of the LOC or ROC signal. Only merged REM were considered as gold standard REM, in this way, at least 2 raters must agree on a REM movement before it is marked as such. Multi peaked waves (as marked by a single rater) closer than 120msecs did not occur, REMs further apart were always considered as separate movements.

Feature Extraction

Features were extracted from each consecutive window of filtered LOC, ROC and the negative product of LOC and ROC (NEGP) (as proposed in Agarwal et al, 2005, this signal is maximal during REM movements). These features can be broken into 4 broad categories: time-domain features, frequency domain features, time-frequency domain features, and nonlinear features. This does not suggest that all time, frequency and time-frequency domain techniques are linear but rather they are not based on the theories of dynamical systems. Features were generally based on apriori theoretical intuition (e.g. REMs have higher amplitudes, thus amplitude is considered a good feature). However, human scoring is a subjective process that cannot be perfectly captured by a discrete set of rules, thus a range of features was implemented.

Feature Reduction

A backward elimination stepwise algorithm was used to reduce the feature space. Here, the classifier starts with the set of all candidate features, then removed each feature separately while measureing algorithm performance (F1 Score). If removal of a feature corresponded to performance improvement, then that feature was eliminated from the set. This process was repeated until performance no longer increased.

Classifier Methods and ECOC

With a window size of 1s, and the high probability of short inter-REM intervals, 2 REMs were often present, therefore, a single Adaboost (binary) classifier, which could only label windows as

containing no REM (0) or REM (1) would underestimate REMs. To limit classifier confusion to rare cases, windows containing 3 REMs in the gold standard were set to 2 REMs (<0.1% of windows), yielding 3 classes to distinguish between windows containing No REM (90%), 1 REM (8%), 2 REM (2%).

Error-Correction Output Codes (ECOC) was used to overcome this 3-class problem. ECOCs provided a method of coding and decoding multiple one vs one binary classification decisions into multi-class decisions. With three classes to classify, three separate dichotomous Adaboost classifiers are required, where each classifier learns to split a pair of classes (outputting a -1 or 1) while ignoring the third (classifier is not trained on the 3rd class and will output an erroneous -1 or 1). Each window's three classifier outputs can be compared via a distance measurement (Laplacian distance gave the best results in this case), to the expected outcome for each class and the closest class is then selected for that window.

As each subject's REMs have unique eccentricities, the importance of algorithm generalizability to new data (other subjects) cannot be understated. Along with internal algorithm cross-validation, k-fold cross-validation was used. Here, the algorithm is trained on 4 of the 5 subjects and then the remaining subject can be used to test algorithm performance. Performance statistics can then be averaged across all 5 different combinations of subjects, with each subject serving as the test subject one time. In this way, the algorithm is never trained on all types of REMs or all subjects, thereby giving greater confidence of generalizability. Within-subject performance, where training and testing data (70:30 ratio) was taken from the same subject is also reported.

Performance Statistics

Specificity and recall, (defined below) are common measures of quantifying algorithm performance. However, in my dataset, the true percentage of windows containing REMs as scored by humans was 6%. With this relatively low ratio of true positives to true negatives, traditional measures of

algorithm performance, such as specificity, are biased (even detecting no REMs give specificity greater than 90%). To overcome bias, I used precision, defined as the percentage of true positives (as determined by the gold-standard) detected by the algorithm. Recall is the number of true REMs correctly classified as such. I use F1 score as a single measure of performance useful in tuning and ranking algorithm performance:

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad F1 = 2\times\frac{Recall\times Precision}{Recall+Precision}$$

The final algorithm makes decisions on the presence of REM in each 1-second window but does not mark exact REM locations. Hence, for algorithm comparison, a windowed version of the gold standard is created by counting the number of gold standard REMs that fall in each consecutive window. Performance statistics are then based on the difference between the REMs per window as classified by my algorithm, and the windowed gold standard. For example, if a window contains 2 gold standard REMs, and the algorithm detects 1 REM, then I have one True Positive and one False Negative. For algorithms that detected individual REM locations (such as my thresholding method), I created a windowed output in the same way as the gold standard.

Additionally, for location-based algorithms, I compared results to the gold standard by marking true positives if a gold standard REM and an algorithm detected REM occur within 200ms of each other, again using MATLABs clustering algorithm. Lone REM's in the gold standard were False Negatives, and lone REM's in the algorithm output were False Positives.

To measure expert and non-expert rater reliability common methods of Cronbach alpha, and inter-rater-agreement are used. Also reported is the average precision and recall of each rater against the gold standard. This precision and recall will be artificially inflated when the gold standard contains that rater (similar to a correlation with itself), therefore I compared each rater to a gold standard created with that rater removed.

Comparison to Previous Work

The finding that REMs may be linked to dreaming led to a flurry of rule-based classifiers implemented with analog electronics (Minard & Krausman, 1971, McPartland et al, 1973, Ktonas & Smith, 1978). As computing power has improved, REM detectors using more complex combinations of features (Agarwal et al., 2005; Doman et al., 1995; Tsuji, Satoh, Itoh, Sekiguchi, & Nagasawa, 2000), matched filtering (Hatzilabrou et al., 1994) and autoregressive modeling (Shokrollahi et al., 2009) have emerged.

These algorithms have impressive performance, but the comparison is biased by differing datasets (EEG channels used, subjects and numbers of human raters used to test performance) and further limited by different performance metrics. Therefore, along with developing my own algorithm, I implemented reported methods from all published LOC and ROC based detectors on my dataset. Note that my aim was to investigate the important features and principles of all the successful published methodologies and provide fair comparison such that future sleep researchers may be better equipped in choosing the right detection algorithm. Detectors using channels other than LOC and ROC (such as vEOG/hEOG) have different signal characteristics and while their methods have been taken into consideration, they were considered out of scope. Table 1 outlines each method and its results. Training and testing with k-fold cross-validation were used when algorithms required tuning, or when exact thresholds were not reported.

Results

Human Scorer Performance

As expected, the expert group raters had a stronger agreement than that of non-experts (Inter-rater-agreement: Expert=0.86, Non-Expert=0.73). This was confirmed by Cronbach Alpha and Precision/Recall for a single rater vs each group (Table 1).

Table 1: Agreement statistics between human raters in the expert group, non-expert group, and combined.

Rater Experience	*Inter-Rater-Agreement*	*Cronbach Alpha*	*Mean Recall (SD)*	*Mean Precision (SD)*	*F1*	*Windowed Gold Standard (GS) used to calculate precision and recall.*
Expert	0.86	0.80	79% (2%)	91% (1%)	.84	Average of each expert compared to Expert GS (with compared expert removed).
Non-Expert	0.73	0.65	68% (15%)	77% (7%)	.72	Average of each non-expert compared to Expert GS.
Combined Raters	0.74	0.68	76% (13%)	83% (9%)	.79	Average of each rater (non-expert and expert) compared to Expert GS (removing compared expert from GS).

Classification Approach Performance

The best algorithm performance averaged across subjects was 78.1% recall and 82.6% precision. The optimum feature set for the machine learning algorithm consisted of Amplitude, Width, Prominence, Rise and Fall Slope, Linear Variance, Cross-Correlation, DWT, DTW, Coastline, Nonlinear Energy, Spectral Skew and Kurtosis, DSI, FSI, and BBDI. The average within-subject performance was 74% recall (SD=9%) and 80% precision (SD=8%), with the best performance for a single subject at 78% recall and 90% precision. Algorithm performance was comparable to that of a single expert human (79% Recall, 91% Precision), and surpassed the average performance of the combined expert and non-expert set (76% Recall, 83% Precision). Considering the mixed experience of technicians in sleep research, the combined group is perhaps a more valid comparison.

Thresholding Approach Performance

Using an intersection combination of extracted features (amplitude, slope, cross-correlation, and Discrete Wavelet Transform), a threshold algorithm reached a performance level of 65.2% Recall and 74.7% Precision. The single best feature of the thresholding approach was peak amplitude in at 75.5% Recall and 59.3% Precision (Figure 8).

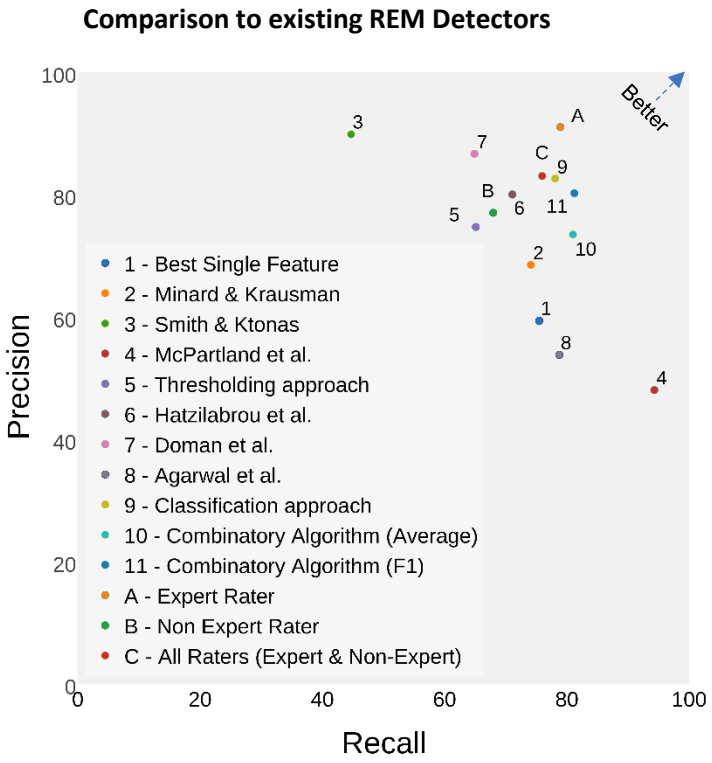

Figure 8: Comparison of REM detector algorithms by Recall and Precision.

My classifier-based method outperforms all others on this dataset (see Figure 8 for a graphical representation of this data). Haztilabrou et al. had the highest performance of the implemented methodologies of past literature (71.1% Recall, 80.0% Precision).

Discussion and Conclusion

Our top-performing algorithm extracts over 25 features from bandpass filtered [0.3Hz-5Hz] LOC and ROC EEG data, and then uses ECOC classifier to train the algorithm to predict REMs from their statistical regularities. The optimum feature set consisted of Amplitude, Width, Prominence, Rise and Fall Slope, Linear Variance, Cross-Correlation, DWT, DTW, Coastline, Nonlinear Energy, Spectral Skew and Kurtosis, DSI, FSI, and BBDI. The automatic detection algorithm presented here is a viable and efficient method of REM detection as it reliably matches the performance of expert human sleep scorers.

By using the same features and algorithm parameters presented here, other researchers can be assured that their definition of a REM movement is consistent with my own. A major advantage of the learning algorithm used is its ability to learn. While my version was trained to match three trained sleep technicians, the learning aspect of the algorithm allows it to adapt to other expert gold standards. Thus, while performing well 'out of the box', the algorithm could also be automatically tuned to suit different detection needs. Furthermore, the algorithm could potentially be adapted to detect other features of sleep, such as sleep spindles. It is important to note that if researchers believe that REM movements in their population are significantly different from those presented here, or their own definition of REM is significantly different, feature sets and thresholds will no longer be optimal and performance will decrease. In this case, it is advised that algorithm retraining is undertaken. To quantify this difference researchers can mark REM movement peaks and then create statistical distributions (MATLAB code provided).

I feel my dataset is of sufficient size for algorithm training, however, as with all machine learning algorithms, it is preferable to have more data to increase generalizability (Domingos, 2012). Current work by the authors includes creating a massive, open, online sleep dataset with expert annotations of REM, spindles and other sleep features to allow for better algorithm training and validation.

Importantly, any algorithm will only be as precise as its gold-standard. My data show that the agreement between my experts is not unanimous and the disparity between experts and non-experts shows some level of learning required for expertise. Since the classifier algorithm learns from humans, it is inherently limited by the agreement between observers. To achieve optimum performance, the validity of the gold-standard must increase. While adding more raters (and hence reliability) does not necessarily mean increased validity, it does create a more reliable and generalizable standard. Since each sleep lab will have different criteria for scoring REM events, more algorithm generalizability is preferable. Gathering scored REM sleep data from expert sleep scorers across many different labs is

possible, or potentially raters could be crowdsourced. Warby et al. (2014) used crowdsourcing technology to create a much larger pool of expert (24) and non-expert (114) raters for sleep spindle algorithm comparison (Each epoch was viewed approximately 5 times by experts and 10.7 times by non-experts). They also found that adding 3 confidence levels (1= not a spindle, 2 = unsure, 3= definitely a spindle) to spindle scoring lead to a better gold-standard and more agreement among raters. Similar methods would benefit future REM scoring.

A limiting factor of the current machine learning algorithm is its inability to directly pinpoint REM locations. The resolution is limited to a 1-second window. However, the current algorithm is suitable for the investigation of REM density. If exact locations are required, then the thresholding algorithm or past algorithm implantations can be used. While each algorithm employed appropriate cross-validation techniques to reduce overfitting, when selecting the best algorithm from a set of algorithms (we estimate approximately 100 variants tested), the choice is dependent on test data. This form of overfitting, common to algorithm development, where one algorithm may be chosen over another because it happened to perform well on this particular dataset, may affect my results and impact generalizability.

By reducing researcher time and effort, algorithms to detect NREM sleep spindles, have begun to give insight into the role of sleep EEG features in cognition (Mednick et al., 2013; Wamsley et al., 2012). Similarly, considering the strong link between REM sleep and memory (Genzel, Spoormaker, Konrad, & Dresler, 2015), the literature on the role of REMs in cognition is remarkably sparse. Thus, the use of an automatic, reliable and time-saving detector may increase the number of research studies addressing this issue. In this view, my versatile REM detector adds an additional piece to the sleep researcher's toolbox and aids the quest to understand the role of rapid eye movements in biology and cognition.

Chapter Four:

P(Sleep) 2.0 - A continuous model of sleep architecture and features

The follow chapter introduces a mathematical model of sleep known as *"P(Sleep) 2.0",* and combines previous work of the author (a model of sleep architecture known as *"P(Sleep) 1.0",* and an automatic sleep feature detector, (Yetton et al., 2016)) to develop a continuous, probabilistic model of the dynamics of sleep stages and sleep features across a night of healthy sleep. This model quantifies and predicts sleep stages and features at an epoch by epoch timescale. Parameters of this model are analyzed to gain a deeper understanding of sleep dynamics and its influences. This work represents a definitive model of sleep dynamics, both at the macros structure level (i.e. sleep stages) and the microstructure (EEG sleep features). It acts as a central reference for the patterns of healthy sleep and thus informs research, clinical and public health endeavors.

Methods

Data sources

The datasets proposed for analysis include 12 investigated in Yetton et al, 2018: 3 datasets from the Furman Sleep Lab (Wamsley, Hamilton, Graveline, Manceor, & Parr, 2016), 4 open-source sleep datasets available on the National Sleep Research Resource (NSRR)(Blackwell et al., 2011; Blank et al., 2005; Dean et al., 2016; Hibbs et al., 2014; Orwoll et al., 2005; Quan et al., 1997; Redline et al., 1998; Resnick et al., 2003; Spilsbury et al., 2005), and 5 from the Montreal Archive of Sleep Studies(O'Reilly et al., 2014) (MASS). A new dataset included was HomePAP from the NSRR for a total of 13 datasets.

Data Inclusion Criteria

Subjects were considered for inclusion if they were between 18 and 90 years old, had their first bout of sleep between 10 pm and 2 am, and total sleep time (including WASO) of 6hrs or greater. Subjects were excluded if they reported high alcohol consumption, previous stroke or heart attack or a

previous diagnosis of Parkinson's, Alzheimer's, depression, narcolepsy, restless leg syndrome, excessive sleepiness or insomnia. When available, apnea and hypopnea indexes were used to exclude subjects (excluded if Apnea-Hypopnea Index > 10). Additionally, if subjects reported taking psych medications currently or in the past, or had taken sleep medications in the last 3 days they were removed. To exclude non-normal sleepers an age varying threshold was placed on sleep efficiency. This followed the sleep efficiency of "normal sleep" quantified by (Ohayon et al., 2003) where subjects with sleep efficiency below a linear line from 85% sleep efficiency at 20 years old, to 65% sleep efficiency at 90 years old were excluded. After exclusion criteria were applied, there were 1003 nights of sleep remaining for modeling. Demographic variables are reported in Figure 9.

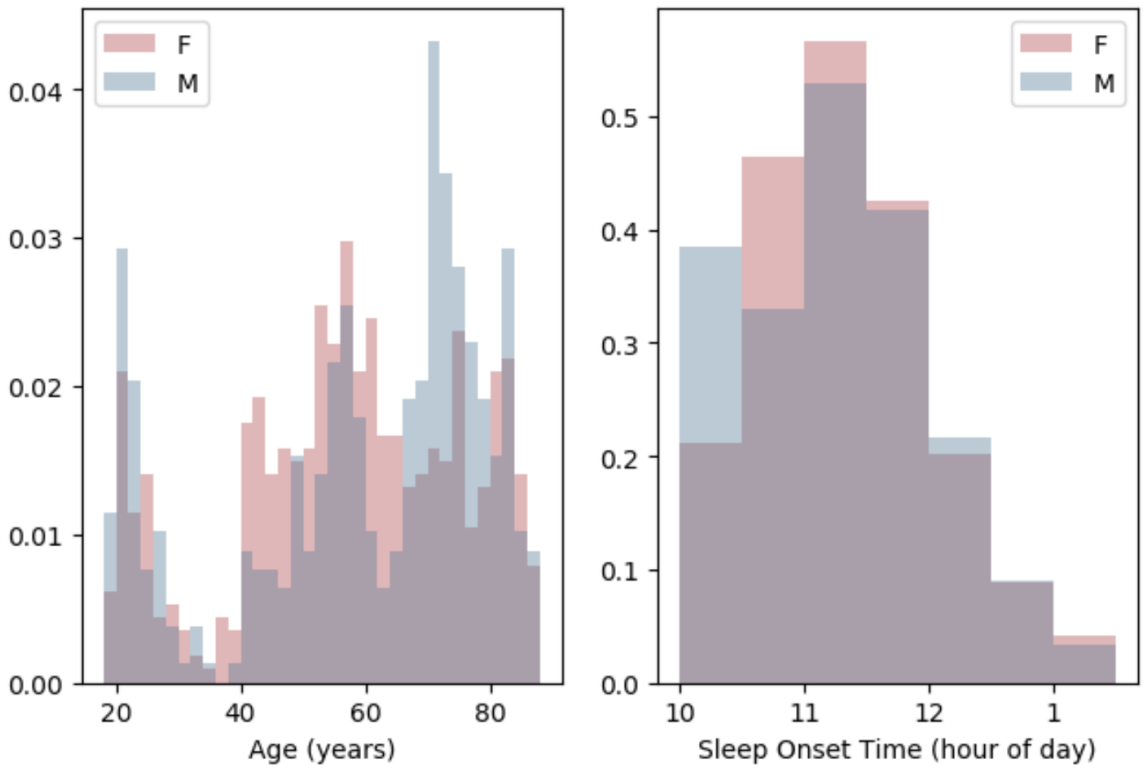

Figure 9: Descriptive statistics of the full dataset. Left: distribution of age measured in years, split by sex. Right: Distributions of sleep onset time (initial epoch of sleep) split by sex. Values greater than 10 represent PM, values lower than 2 are AM. Data was limited to subjects between 18 and 90 and with sleep onset times between 10 pm and 2 am.

Data Engineering Pipeline

With the large quantity of data (multiple terabytes of PSG), the number of variables, and heavy preprocessing required for sleep feature extraction, the importance of data engineering cannot be understated. I, therefore, developed a full-stack application, the MednickDB, for sleep data management (grouping data for a specific subject together), automatic parsing (extracting sleep features), and data query (filtering data on specific attributes). This project represents a considerable undertaking, spanning over 10,000 lines of code and without it, this project would not have been possible. While the full technical specifications of the MednickDB are out of scope, the basics are discussed here due to its critical nature in enabling this project (Figure 10):

1. Data files (PSG, demographic information, scorefiles) are input to the MednickDB via a python API. Metadata such as study and subject information is also included. Supported filetypes include most formats of EEG (.eeg, .edf), sleep scoring (Hume, Grass, XML, EDF), and other tabular data (e.g. CSV or excel files containing demographics or health data).
2. Files are stored according to study, subject and visit metadata on a large virtual machine hosted by the UCI Social Science IT team.
3. Newly added files are parsed by python microservices, which run sleep feature extraction algorithms. Algorithms produce several sleep feature variables per sleep epoch (such as counts, durations, amplitudes, etc). All microservices had associated unit tests (automated test scripts) to reduce the possibility of coding/algorithmic errors.
4. Data extracted from these algorithms are stored per study, subject and visit.
5. Data from healthy subjects was then downloaded via the python API and pre-processed to a modellable format (see *Modeling* section)

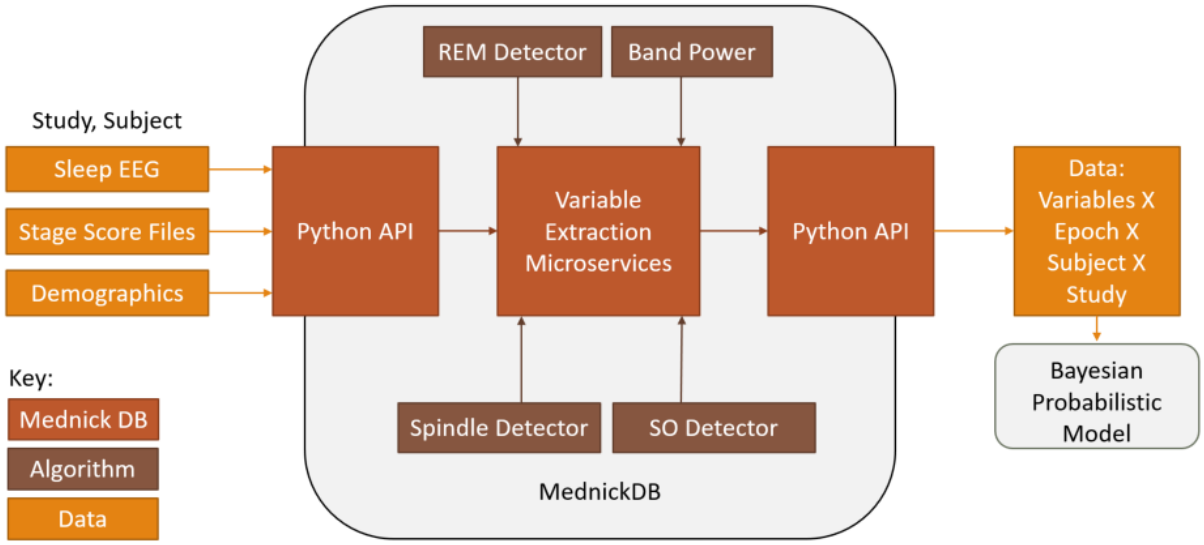

Figure 10: Overview of the MednickDB data processing pipeline.

Data Preprocessing

Raw data were processed by the MednickDB microservices:

Sleep scoring

Sleep scoring files were standardized into 30-second epochs, measured from lights out. Bedtime was assumed to coincide with lights off for all records, and the time between lights off and the first epoch of sleep was considered sleep latency. Durations and sleep features of wake prior to the first epoch of sleep (wake before sleep onset - WBSO) was not analyzed, but transitions from WBSO to the first epoch of sleep were modeled. Wake between the first and last epoch of sleep was considered WASO, and any after the last epoch of sleep was not analyzed.

EEG Standardization and artifact detection

All EEG/EOG was resampled to 256Hz and bandpass filtered between 0.3 and 35Hz. Voltages were standardized to μV. Channels were reordered, given standard names and referenced to linked mastoids. Epochs with artifacts were removed following the procedure in Purcell et al: Each epoch was compared to a local average of the 7 epochs before and after it. An epoch was excluded if the beta or

delta power was 3 times the local average. Further, metrics of root mean square and three Hjorth parameters (activity, mobility, and complexity) were extracted, and epochs were excluded if they were +- 2.s.d from the mean across all epochs for any of these measures. Approximately 7% of epochs contained artifacts, much of which was WASO. EEG features of spindles and slow oscillations were extracted from the C3 channel only (and results only generalize to this channel). REM events were extracted from LOC/ROC. Records missing C3 or LOC/ROC were excluded from band power or sleep feature analysis (N=104).

Spindles Extraction

There are many available spindle detection algorithms, each with various biases. There is no measure of a true spindle, however, myself and colleagues have recently released a large, high-quality dataset of spindles, which collates the spindles of many human experts. Two spindle algorithms achieve a sufficiently high match with human scoring Lacourse et al. (2018) and Wamsley et al. (2012). Due to the long runtime of the Lacourse et al. algorithm, and the large N in this study, I opted for Walmsley et al. algorithm. This algorithm operates on artifact-free NREM only, and first uses an 8 parameter Morlet wavelet to transform the signal into the frequency band, which is then squared. The complex part of the signal is discarded, and the real part is squared. A moving average is created using a 0.1-second sliding window. A spindle event is then identified whenever the moving average signal exceeds 4.5 the mean signal amplitude. Spindles with a duration of fewer than 0.4 seconds were rejected.

Slow Oscillations Extraction

The Wonambi package's implementation of Massimini (2004) slow oscillation detector was used. Parameters were updated to those defined by the AASM manual (Berry et al., 2012). Only artifact-free NREM stages were inputted to the algorithm. Signals were then band passed between 0.1 and 4 using a 4^{th} order Butterworth filter. Slow waves are then detected if they had 1) a positive zero crossing

separated by at least 0.25 seconds, but not more than 1 seconds, 2) a negative peak between the two zero crossings with voltage less than -40 µV, and 3) a peak-to-peak amplitude greater than 75 µV.

Rapid Eye Movements Extraction

Hatzilabrou et al. (1994) Rapid Eye Detection algorithm was used, as implemented in Yetton et al. (2016). While the machine learning algorithm from Yetton et al has higher accuracy, it is time-consuming and was not feasible to run given the size of this dataset. Only artifact-free REM epochs were analyzed. Details of this algorithm are in chapter 2. The counts per epoch of sleep features are reported in Figure 11.

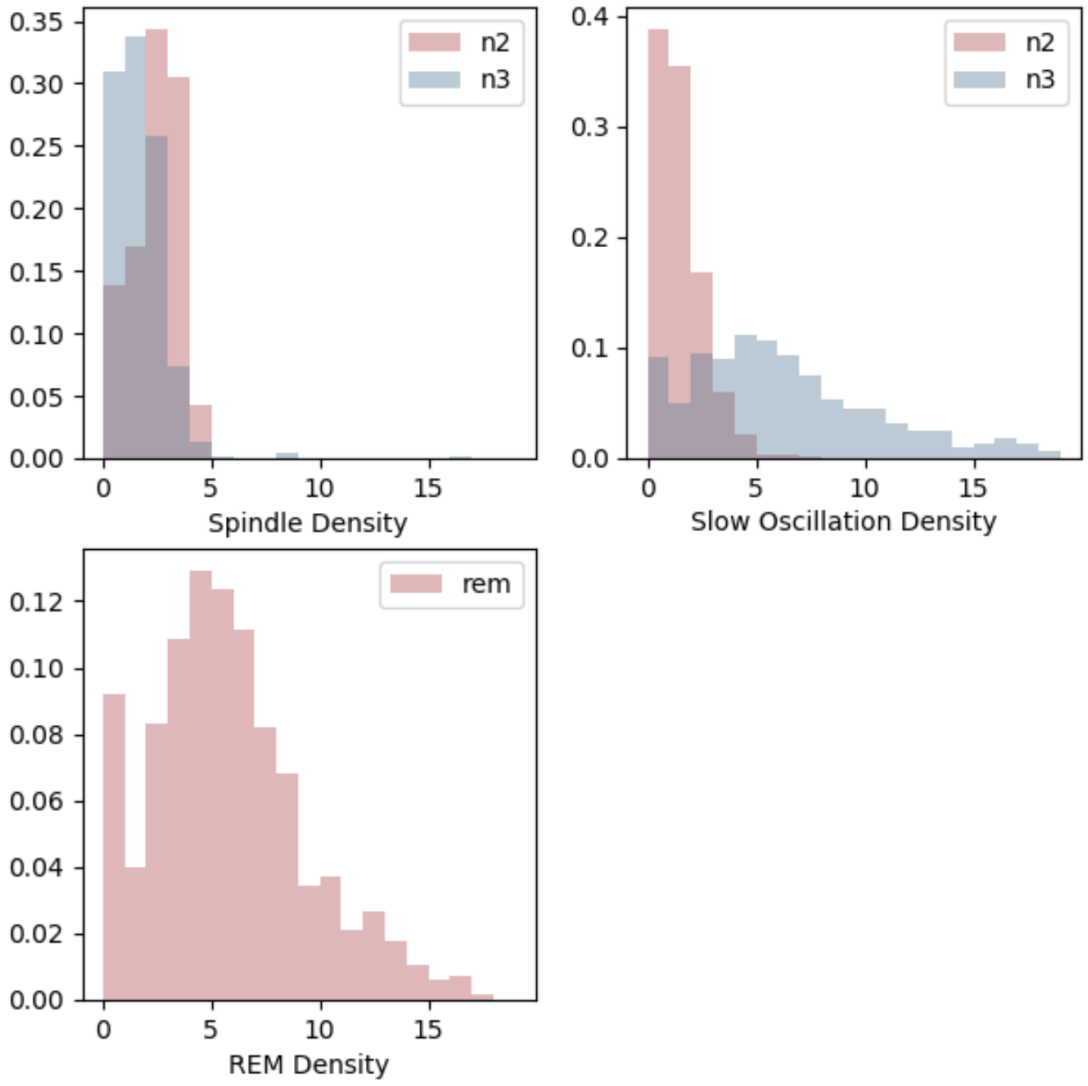

Figure 11: Distributions of density (counts/minute) for Spindles, Slow Oscillation and REM events for each sleep stage considered separately. Y-axis is probability density.

Modeling

We consider 3 models types, each with a different set of outcomes:

1) Models of sleep dynamics (predicting the next sleep stage).
2) Models of Spindles and Slow Oscillation count.
3) Models of REM event count.

While there are likely correlations between these sets of outcomes, due to computational infeasibility, this was not considered. Correlation within the set of outcomes (i.e. between spindles and SO) was considered, however (see modeling paradigm).

We included the following variables as predictors:

- Stage of the current epoch: One of WASO, n1, n2, n3 or REM.
- Stage of the previous bout: The sleep stage prior to transitioning to the current sleep stage (includes the above stages plus WBSO).
- Tau: The time in the current stage (measured in minutes since transitioning to this stage)
- Time of day: Clock time, and coded as hours since 10 pm.
- Age: Age was normalized when fitting models, but appears as raw age in plots to ease interpretability.
- Sex: Biological sex (Male/Female).

All models consider a single data-point as a single 30-second epoch of sleep. Each data-point, therefore, contained the above variables for each epoch. Note that age and sex are at the “per subject” level rather than the “per epoch” level, but this nesting was not considered. An example of the temporal data used

to train the model is shown in Figure 12. Depending on the model trained, some of this data was ignored (e.g. sleep stage models do not include sleep feature counts as predictors).

Sub	**1**	1	1	1	1	1	1	Per Sub (s) | Study
Age	**22**	22	22	22	22	22	22	
Sex	**F**	F	F	F	F	F	F	
$Stage_{e+1}$	W	1	1	2	2	2	2	Per Epoch (e)
$Stage_{e}$	W	W	1	1	2	2	2	
$Duration_{b}$	30	60	30	60	30	60	90	Per Bout (b)
$Stage_{b-1}$	W	W	W	W	1	1	1	
Time of Day	23:00:00	23:00:30	23:01:00	23:01:30	23:02:00	23:02:30	23:03:00	Per Epoch (e)
Spindles	0	0	0	1	2	4	2	
SO	0	0	0	1	4	2	1	
REM	0	0	0	0	0	0	0	

Figure 12: Example of data for each variable (row) for each epoch (columns). e=current epoch, b=current bout. W=wake, F=Female, SO=slow oscillation.

Model Definitions

Sleep Stage Dynamics

In my previous work, I used low-resolution categorical distributions to model the duration of the next stage. To increase temporal resolution, P(Sleep) 2.0 models sleep at an epoch by epoch timescale. Therefore, instead of predicting a series of {stage, duration} bouts (as in *P(Sleep) 1.0*), I allowed stages to transition to themselves, where transitions occur at every epoch (i.e. a Markov chain). A Markov chain models a sequence of discrete stages, where the next stage in a sequence is categorically sampled from a row of a N x N matrix of transition probabilities. Here, a specific row is indexed by the current stage to give an array of probabilities for the next stage:

$$\theta_{base[Stage_{current},\ Stage_{next}]} \in \mathbb{R}^{5\times 5}$$

$$\theta_{base} \sim Normal(\mu = 0, \sigma = 100)$$

$$\theta = \theta_{base[Stage_{current}, :]}$$

$$stage_{next} \sim Categorical(Softmax(\theta))$$

This is was the simplest model of sleep dynamics, where transition probabilities are simply modeled as constant across time and demographics ($\theta = \theta_{base}$). This model represented the "baseline" for which to compare more complex models. Elements of θ_{base} are sampled from a normal distribution, and rows of θ were softmax transformed so they add to one (i.e. so they represent probabilities of the next stage). Bout duration will be implicitly captured in the sequence of stages (e.g. [N1, N1, N1, N1]): the longer the duration of a stage is, the higher the self-transition probability will be (i.e. diagonals of θ). For example, N1→N1 ($i.e.\ \theta_{[1,1]}$) will have a lower probability than N3->N3 ($i.e.\ \theta_{[3,3]}$) because N3 bouts generally last longer.

Considering the nested structure of the data

Each dataset contains data collected with different equipment and scored by different PSG technologists. Therefore, data may not be independent and identically distributed if *dataset* is not considered as a grouping variable. As the first step in each of the 4 models types, I test if a hierarchical model, where dataset is used as a grouping variable, fits better (lower LOO) compared to the simpler pooled model, where all data modeled as identically distributed. I modeled the effect of dataset on the baseline transition probabilities only (i.e. a random intercept model). To alleviate divergences caused by the hierarchical funnel (Betancourt, Girolami, & Carlo, 2013), I followed a non-centered parametrization of the dataset groups:

$$\theta_{base\_means} \sim Normal(\mu = 0, \sigma = 100)$$

$$\sigma_{dataset\_varience} \sim HalfNormal(\sigma = 100)$$

$$\theta_{base\_per\_dataset} \sim \theta_{base\_means} + \sigma_{dataset\_varience} \cdot Normal(\mu = 0, \sigma = 1)$$

$$\theta_{base} = \theta_{base\_per\_dataset[Stage_{current}, :, dataset]}$$

I wanted the model to generalize to out of sample datasets (i.e. the model should yield realistic predictions for new data, from new studies, and not just the currently included studies). Therefore, when using the model for prediction and parameter analysis (i.e. drawing conclusions from the models), $\sigma_{dataset\_varience}$ was set to 0. This has the effect of sampling parameters from θ_{base_means} only, and therefore gives predictions that hold across all datasets, and hence generalize to new, out of sample data.

Adding predictors to the sleep stage model

Many other variable influence the pattern of sleep. I encode their influence as linear additions to θ. From *P(Sleep) 1.0*, we know that the identity of the previous *bout* influences the probability of the next stage. The effect of the previous bout was included by adding (or subtracting) a small probability to each element of θ, where the amount added changed based on the stage of the previous bout. More concretely, the effect of the previous bout is encoded as a 3-dimensional matrix, with dimensions of *stage of the previous bout*, *stage of the current epoch*, and *stage of the next epoch,* and I index into this matrix using the stage of the previous bout and previous epoch, then add this to θ. For example, there are 5 possible sleep stages, then $\theta_{previous} \in \mathbb{R}^{6\times5\times5}$, where dimensions are previous bout, current stage, and next stage respectively. The transition probabilities then become (Figure 13):

$$\theta_{previous} \sim Normal(\mu = 0, \sigma = 100)$$

$$\theta = \theta_{base[stage_{current},\ :]} + \theta_{previous[bout_{previous},\ stage_{current},:]})$$

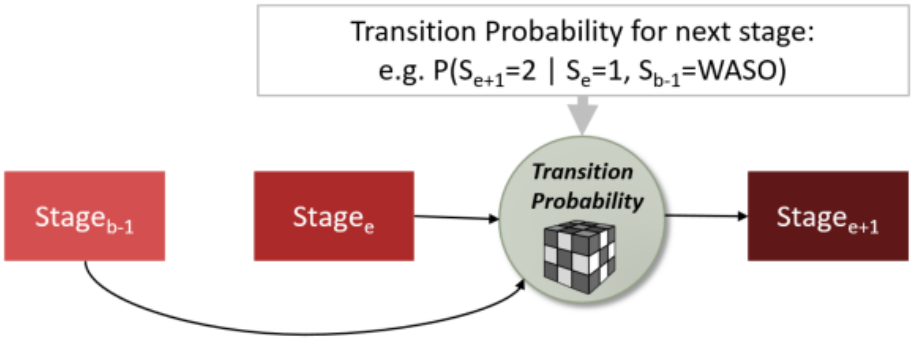

Figure 13: Predicting the next stage as a Markov chain, where the previous two stages (S_e, S_{b-1}) influence the probability of the next stage (S_{e+1}). Note that the previous stage may include WBSO: wake before sleep onset, and therefore has 6 possible values.

Not all stages durations follow the same geometric functional form enforced by a standard Markov chain (which leads to exponential bout durations). I, therefore, included the effect of the duration of the current stage (how long we have been in the current stage for, *tau* - τ) as inputs to a set of functions $\theta_{f(\tau)}$ that further modify transition probabilities (Figure 14). Previous literature has suggested that the duration of sleep stages follows specific distributions (such as modified exponential, power law, etc)(Bizzotto et al., 2010; Chu-Shore et al., 2010; Kishi et al., 2018). The branch of statistics known as survival analysis attempts to model how long some process, human, widget – or sleep stage – survives. Literature from this field helped convert previously suggested functions of stage duration into *hazard functions,* which give the probability of something ending (i.e. transitioning or not) in the next time step. By inputting tau into these hazard functions, I output a probability that further modulates transition probabilities. A number of hazard functions based on the literature were trialed (Pareto – i.e. power distribution, exponential, gamma, and Weibull - similar to stretched exponential). I considered a different set of hazard function parameters for each stage (i.e. hazard functions are dependent on $\text{stage}_{\text{current}}$. Elements of θ are now dependent on the current stage, its duration (τ), and the stage of the previous bout.

$$\theta = \theta_{base[stage_{current}, :]} + \theta_{previous[bout_{previous}, stage_{current}, :]}) + f_{hazard}(\tau, stage_{current})$$

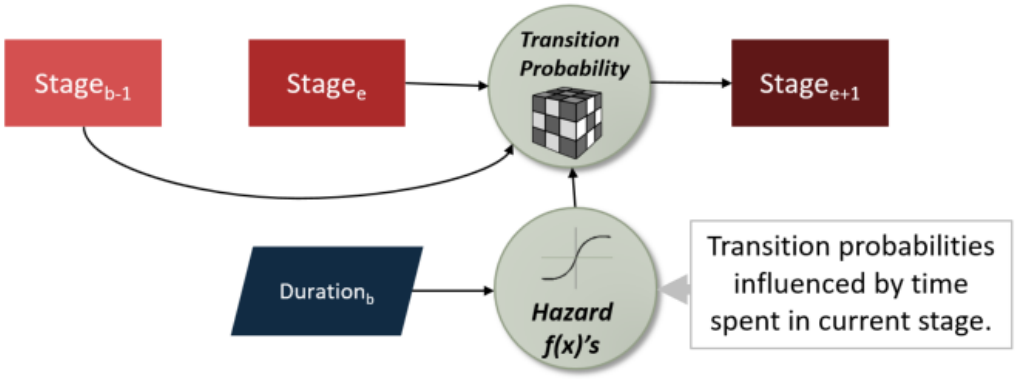

Figure 14: Duration effect on transition probabilities

All other variables influenced the next epoch as additional terms of θ, and parameters varied based on the current stage.

1) A simple linear effect:

$$\beta = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$
$$\theta_{variable[stage_{current},\ stage_{current}]} = variable \cdot \beta_{[stage_{current}]}$$

2) A quadratic effect:

$$\beta_{linear} = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$
$$\beta_{quadratic} = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$
$$\theta_{variable[stage_{current},\ stage_{current}]} = variable \cdot \beta_{linear[stage_{current}]} + variable^2 \cdot \beta_{quadratic[stage_{current}]}$$

3) An exponential effect:

$$\beta_{lamda} = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$
$$\beta_{alpha} = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$
$$\theta_{variable[stage_{current},\ stage_{current}]} = \beta_{alpha[stage_{current}]} e^{variable \cdot \beta_{lambda[stage_{current}]}}$$

4) A hyperbolic effect (for sleep stages only):

$$\beta_{alpha} = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$
$$\theta_{variable[stage_{current},\ stage_{current}]} = variable / \beta_{alpha[stage_{current}]}$$

5) A cyclic (sinusoidal) process, for time of day only, borrowed from the two-process model:

$$\beta_{phase} = VonMiess(\kappa = 0.01) \in \mathbb{R}^5$$
$$\beta_{amp} = Normal(\mu = 0, \sigma = 100) \in \mathbb{R}^5$$

$$\theta_{variable[stage_{current},\ stage_{current}]} = \beta_{amp[stage_{current}]} \cdot sine(2\pi \frac{variable}{24} + \beta_{phase[stage_{current}]})$$

The complete model's theta was a sum of a base-rate effect, previous bout effect, time in stage (tau) effect, and other effects (one of the above 5 terms for each variable considered):

$$\theta = \theta_{base[stage_{current},\ :]} + \theta_{previous[bout_{previous},\ stage_{current},:]}) + f_{hazard}(\tau, stage_{current}) + \theta_{variable1[stage_{current},\ stage_{current}]} + \cdots + \theta_{variableN[stage_{current},\ stage_{current}]}$$

Modeling other outcomes:

Models REM events, and SO and spindle events were similar to models of the next epoch. Theta was formed the same way by first adding a baseline effect (e.g. across all epochs, what is the average amount of spindles), then the effect of the previous stage, and then adding the effect of other variables as either linear, quadratic, hyperbolic, exponential or sine. How the outcome was sampled from theta (and therefore its interpretation) changed between sleep stage and sleep feature models:

1. SO and spindles: Only N2 and N3 stages were considered. Spindle and slow oscillations are counts within an epoch. The Poisson distribution was the obvious choice for count data. However, the Poisson rate cannot go negative, so I wrapped it in a softplus function to guarantee all elements were positive. SO and spindles are related, and the number of spindles and SO per epoch may be correlated, therefore the covariance between these rates were included in some models using a similar multivariate normal as band power. Theta can be interpreted as the mean count of a feature during an epoch.

$$\sigma_{band\_varience} \sim HalfNormal(\sigma = 100) \in \mathbb{R}^5$$

$$\sum_{band\_covarience} \sim LKJCholeskyCov\big(\sigma = \sigma_{band\_varience}, \eta = 5\big) \in \mathbb{R}^{5 \times 5}$$

$$\theta \sim MultiVarNormal(\mu = \theta, \Sigma = \sum_{band\_covarience})$$

$$Spindle_{counts}\ and\ SO_{counts} \sim Poisson(\mu = softplus(\theta))$$

2. REMs were found to be exponentially distributed rather than Poisson. Theta is interpreted as the mean REM count (number of REM events per epoch).

$$REM_{counts} \sim Exponential(\lambda = 1/softplus(\theta))$$

Modeling Procedure

Models were implemented in pymc3, a probabilistic modeling package for the Python programming language (Salvatier et al., 2016). The No-U-Turn gradient-based Markov Chain Monte Carlo Sampler (Hoffman & Gelman, 2014) was used to fit each model, with 500 tuning (burn-in) samples, and 2000 parameter draw samples. The Pymc3 package provides diagnostics such as divergences (MCMC sampling has become biased), and the Gelman-Rubin R static, which verifies that two runs of the model fitting procedure give the same (or very similar) parameter posteriors**.** The models presented here have no divergences and Gelman-Rubin statistics of less than < 1.05 for all parameters, which indicated successful fitting procedures for all.

Finding the optimal model proceeded in a stepwise fashion. At each step, I compared a set of models, each encoding a different hypothesis, and whichever model had the lowest LOO (see *Evaluation Criteria* section) was considered the best and formed the baseline for the next set of comparisons. For example, I first compared if the baseline model, or a baseline model where dataset was a grouping variable, was better. Then to the better of the two models, I added the effect of the previous bout. If this improved fit, it became the new baseline. To this new baseline, I added a new model for each parametrization of tau (linear, quadratic, etc), and then compared each to the baseline. This process continued until the model fit did not increase, or there were no more variables to include. The order of variables was: previous bout, tau, time of day, age, age^2, sex, age X sex. Only a subset of the full data (30,000 data points) was used during this procedure to make computational time tractable. These data points were randomly sampled, and therefore represents the same distribution as the original dataset.

After the best model was found, its parameters were re-estimated with the full dataset (a process taking many days).

Interpreting the effects of each parameter

The flexibility in modeling that the MCMC procedure and Bayesian approach affords has major advantages in it can capture complex relationships in the data, however, because models are no longer simple independent combinations of normal distributions, interpretation of model parameters is less straightforward. This is especially true for models of next stage, where parameters are categorical probabilities, and therefore must sum to one (enforced here through a softmax transform). For example, if age has a positive linear effect on N1→N1 transitions and transitions from N1 to any other stage must sum to one, then age necessarily causes a reduction in one (or more) of the other N1 transitions. Continuing this example, as age increases, the transitions probability from N1→N1 also increases, but it is constrained to always be less than one minus the sum of other N1 transitions. As the transition probability approaches this bound, it will asymptote, leading to non-linear effects of age. In the P(Sleep) 2.0 model of the next stage, the parameters that effects transition probabilities are difficult to interpret. Instead, I use the model to predict transition probabilities as a function of one or more variables of interest, while holding other covariates constant (at their mean) or by marginalizing out the effect of these covariates (for categorical variables like the previous stage). Because parameters are distributions and not simply point estimates, I then sample 100 draws of each parameter from the joint posterior distribution (i.e. the MCMC trace) and obtain the prediction for each parameter draw. The mean of these predictions, as well as the standard deviation, are plotted and then interpreted.

Priors

A major advantage of the Bayesian approach is the ability to encode known theory as priors on model parameters. Given the amount of data used, priors generally play an insignificant role in the final posterior distribution of each parameter (the data is able to provide far more information than the

prior). However, priors serve here to initialize the MCMC sampling procedure in regions of high posterior density, thereby speeding sampling, and reducing the chance of divergences (a special case of the No-U-Turn sampler failing) which leads to biased inference. Prior values were gathered from previous literature when available, or uninformed priors were used when no previous literature could be found. Some papers only reported the direction of an effect and not exact parameter values. In this case, I estimated realistic values for priors that captured this directionality, but enforced it weakly (i.e. for a normally distributed parameter, a positive effect garnered a realistic positive mean, but large standard deviation). The previous literature used to estimate priors for each model are shown in Table 1.

Table 2: Literature used for priors

Model	References for priors:
Sleep Stages	(Carskadon & Dement, 2005; Yetton et al., 2018)
Spindles	(Purcell et al., 2016; Warby et al., 2014; Yetton et al., n.d.)
Slow Oscillations	(Carrier et al., 2011; Massimini, 2004)
REM Density	(Ficca et al., 1999; Khalsa, Conroy, Duffy, Czeisler, & DIJK, 2002; Kovács, Kosztolányi, & Kis, 2018; Reynolds III et al., 1985; Yetton et al., 2016)

Evaluation Criteria

Model fit was quantified using Leave One Out (LOO) cross-validation method optimized for MCMC samples (Vehtari, Gelman, & Gabry, 2016). In standard LOO, a model is fit on all samples but one, and the last sample is predicted and compared to the observed variable. The procedure is then replicated such that each sample of the dataset has been predicted. These predictions are compared to observed values, then logged and summed. A reliable estimate of the standard LOO method, which uses Pareto smooth importance sampling and the trace of MCMC samples generating during model fit has been proposed (Vehtari et al., 2016). This metric approximates the true LOO value by calculating metrics on

the trace of MCMC samples used during model fitting and takes significantly less time than fitting the model itself. The LOO metric for a set of models can be compared, and the lowest metric will give the best fitting model, and therefore stronger evidence for the hypothesis it encodes compared to the other models it was tested against.

Results

The dynamics of sleep stages: models of next epoch

I first investigated the dynamic of sleep using models that predicts the next stage from previous stage information and demographics. I began with the simplest model (transition rates are sampled from some constant base-rate), and then using a stepwise procedure, only added terms when they improved model fit (based on the LOO metric). The sequence of models tested is shown in Table 2. For each of the added variables, I also tested *how* the variables influenced transition probabilities (e.g. was the effect of age best captured as a linear trend, or quadratic trend, etc). I found significant variation in the transition rates across datasets, and therefore a hierarchical model, where each data-point is nested in dataset was used.

Table 3: Model fit metrics for stage models

Base rate	Prev Bout	Tau	Time Of Day	Age	Sex	Age X Sex	LOO	ΔLOO	Weight
Fixed	✗	✗	✗	✗	✗	✗	39461.1		≈0
Random	✗	✗	✗	✗	✗	✗	39456.8	-4.3	≈0
Random	✓	✗	✗	✗	✗	✗	37439.9	-2016.9	≈0
Random	✓	Linear	✗	✗	✗	✗	36913.4	-526.5	≈0
Random	✓	Quadratic	✗	✗	✗	✗	36617.8	-295.6	≈0
Random	✓	Pareto	✗	✗	✗	✗	35811.8	-806	≈0
Random	✓	Pareto	Sine	✗	✗	✗	35702.4	-109.4	≈0
Random	✓	Pareto	Linear	✗	✗	✗	35700.3	-2.1	≈0
Random	✓	Pareto	Quadratic	✗	✗	✗	35692.1	-8.2	≈0
Random	✓	Pareto	Exponential	✗	✗	✗	35690.7	-1.4	<0.01
Random	✓	Pareto	Exponential	Quadratic	✗	✗	35683.6	-7.1	<0.01
Random	✓	Pareto	Exponential	Linear	✗	✗	35682.2	-1.4	<0.01
Random	✓	Pareto	Exponential	Linear	✓	✓	35672.7	-9.5	0.01

Random	✓	Pareto	Exponential	Linear	✓	✗	35662.1	-8.8	0.98

A green tick indicates the model included those variables. A red cross indicates it does not. LOO is the leave one out cross-validation score, while ΔLOO is the difference in LOO between consecutive models. The weight column represents an estimate of the probability that the corresponding model is the best fitting model.

The best-fitting model was 98x more likely than the next best model and included the effects of all variables. I reviewed how each variable influenced transition probabilities by using the best fitting model to predict transition probabilities across the full range of each variable (see methods for details).

Time in stage effects

I found transition probabilities were affected by time in stage (Tau), and the best function for this relationship was Pareto. Figure 8 demonstrates how these transition probabilities changed across the time spent in each stage. Focusing on self-transitions first - higher probability of self-transition indicated longer durations for that stage. From Figure 15, REM and N2 had the greatest durations, followed closely by WASO and N3. Self-transition probability for N1 was low, indicating very short stages (generally lasting less than a minute). Because I modeled transition probabilities as a linear combination of a time-constant base rate, and a time-varying hyperbolic effect, the duration of each stage falls somewhere between exponential and paretally distributed: more paretally distributed stages will have a greater initial rise in self-transition probability before reaching their steady-state value, while exponentially distributed stages have a constant transition probability across time. Stages that have a more parentally distributed duration are more likely to have many extremely short durations (i.e. high fragmentation). Pareto effects were strong for N3, WASO, and low for REM and N1 and N2. This finding was in line with previous literature, where distributions of WASO and N3 were found to follow power laws (same as Pareto), while other stages were exponential (Kishi et al., 2018). Perhaps the most intriguing tau/transition probability result was the extremely pronounced N3 Pareto effect: N3 was more likely to transition to N2 after only a single epoch than it was to remain in N3, however, if an N3→N2 transition did not occur, then N3 was likely to continue for a while.

Turning now to the non-self-transition probabilities, WASO was most likely to transition to N1, and N1 was most likely to transition to N2, although transitions from N1 back to WASO, and to REM were also possible. N2 had a very high propensity to remain in N2 but also transitioned to any other stage (with N2→N1) being the most common. N3 almost exclusively transitioned to either itself or N2. REM transitioned to either N1, N2 or WASO at approximately the same rate, and almost never transitioned to N3.

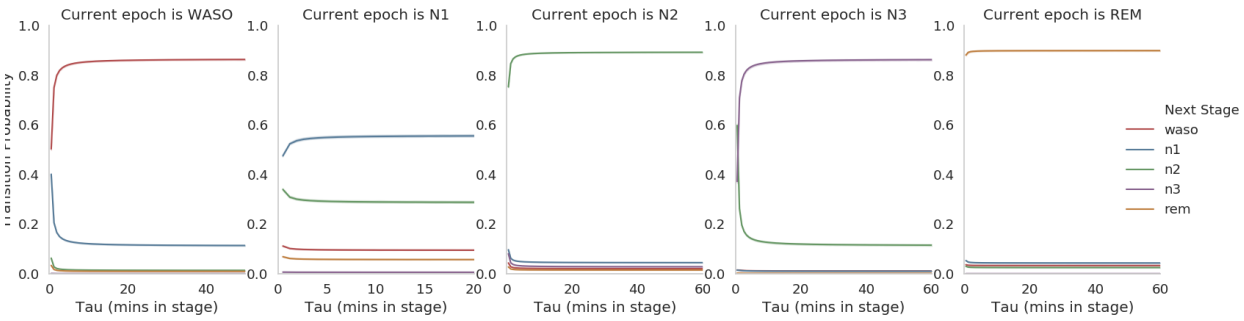

Figure 15: Transitions probabilities from various stages (left to right panels) to various stages (colors of lines) as a function of time in stage (measured in minutes).

Previous Bout Effects

There were strong effects of the previous bout (large change in ΔLOO between a model with, and a model without previous bout). This finding agreed with the previous PSleep 1.0 model, where the previous 2 states are needed to predict the next. In general, the effects of the previous bout were small and distributed across all transitions, however, several interesting findings are included in Table 3, which shows the stage transitions where, by accounting for the previous stage, transition probabilities changed by at least 0.05. For example, if a transition from N2→N1 or REM→N1 occurred, then the next stage was much more likely to be N2 or REM again (i.e. REM and N3 are often broken by N1 bouts, *row 1 & 4, table 3*). Likewise, if an N3→N2 transition occurred, then N3 was more likely to follow (*row 2*, *Table 3*). I also found that when sleep began (the previous stage was WBSO), the duration of N1 bouts were longer (higher N1→N1 probability), and the likelihood of N1-N2 transitions were lower (*row 3 and 6, Table 3*).

Durations of N1 were shorter after an N2 bout (row 8), and, interestingly, the durations of N2 were shorter when following N3 (*row 7 and 8, Table 3*). In the well-characterized general pattern of N3 → N2 → REM → N2 → N3 which occurs across the middle of the night, this later finding means the initial N2 bout (after N3) is shorter than the one following REM.

Table 4: Effect of Previous bout on transition probabilities

Current Stage	**Next Stage**	**Prev Stage**	*Trans P marginalizing over Previous Stage*	*Trans P inc. Previous Stage*	**Difference when inc. previous stage**
N1	N2	N2	0.29	0.45	*0.16*
N2	N3	N3	0.03	0.15	*0.12*
N1	N1	WBSO	0.55	0.66	*0.10*
N1	REM	REM	0.06	0.16	*0.10*
N1	N2	REM	0.29	0.21	*-0.08*
N1	N2	WBSO	0.29	0.21	*-0.08*
N2	N2	N3	0.89	0.78	*-0.11*
N1	N1	N2	0.55	0.44	*-0.12*

Transition probabilities from the current stage to the next stage as a function of the previous stage. The 4th column represents the transition probability when the previous stage is not considered, the 5th columns is when the previous stage is considered, and the 6th column is the difference. Only values where the transition probability changed by more than 0.05 are shown.

Time of day effects

Interestingly, a Sinusoidal relationship of clock-time (time of day) on transition probabilities (as in the Two Process Model's process C) was not observed, instead, an exponential relationship was the best fit (similar to Two Process Model's process S). This model only considered the sleep period, if you modeled sleep stage propensity across a 24 hr period, a sinusoidal effect of time of day model may fit better (and better fit the narrative of the two-process model). The effect of transition probabilities across the time of day was not consistent across stages and is shown in

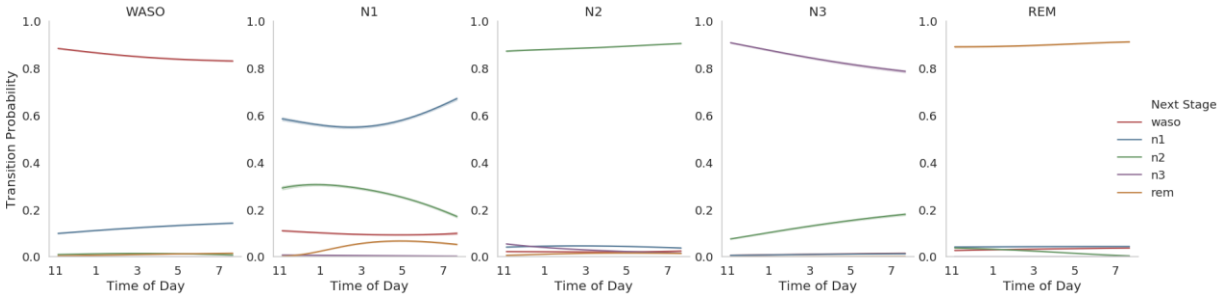

Figure 16. As time of day increased, self-transition probabilities, and therefore durations of WASO and N3 decreased, while N2 and REM increased. A curvilinear pattern was observed for N1, where N1 durations reduced until 3 am, then rose towards morning. For most stages, the change in self-transition across the night was mirrored by an opposite change in the most likely non-self-transition, for example, the reduction in N3→N3 transitions was accounted for by a rise in N3→N2 transitions. For N1, it was noted that N1→REM transitions were extremely unlikely during the first quarter of the night, and then preceded to rise through the second quarter of the night, and remain constant after this. This means that the greater proportion of REM during early morning sleep is a combination of a sharp uptick in N1→REM transitions, and increasingly longer REM bouts (less fragmented REM).

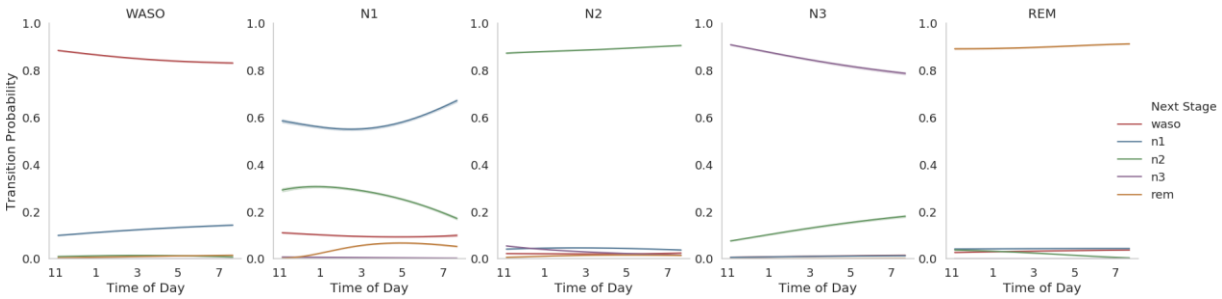

Figure 16: Transitions probabilities from various stages (left to right panels) to various stages (colors of lines) as a function of time of day (measured in hours).

Sex Effects

I found a main effect of sex, but no explicit age*sex interaction. For self-transition probabilities, and therefore the duration of stages, females had longer bouts of WASO, N3, and REM (less fragmentation) but shorter bouts of N1. There was a minimal difference across sex for N2 self-transition probabilities. Males were more likely to transition from WASO directly to REM, from N1 to WASO/N2 and from N2→N3. The only (non-self) transition probability that was higher for females than males was N2 to N1.

Age Effects

While the difference between a linear and a quadratic effect of age was small, a linear effect had a slightly better fit. Note that because there is a different age parameter for each transition, and

that transition probabilities from a stage are normalized to one (through a softmax transform, see methods), there can be minor non-linear effects across age. In general, age effects sizes were small (Figure 17). There was a subtle effect of age on the duration of all stages, with an increase in WASO, and a decrease in the duration of all sleep stages (and therefore more fragmentation).

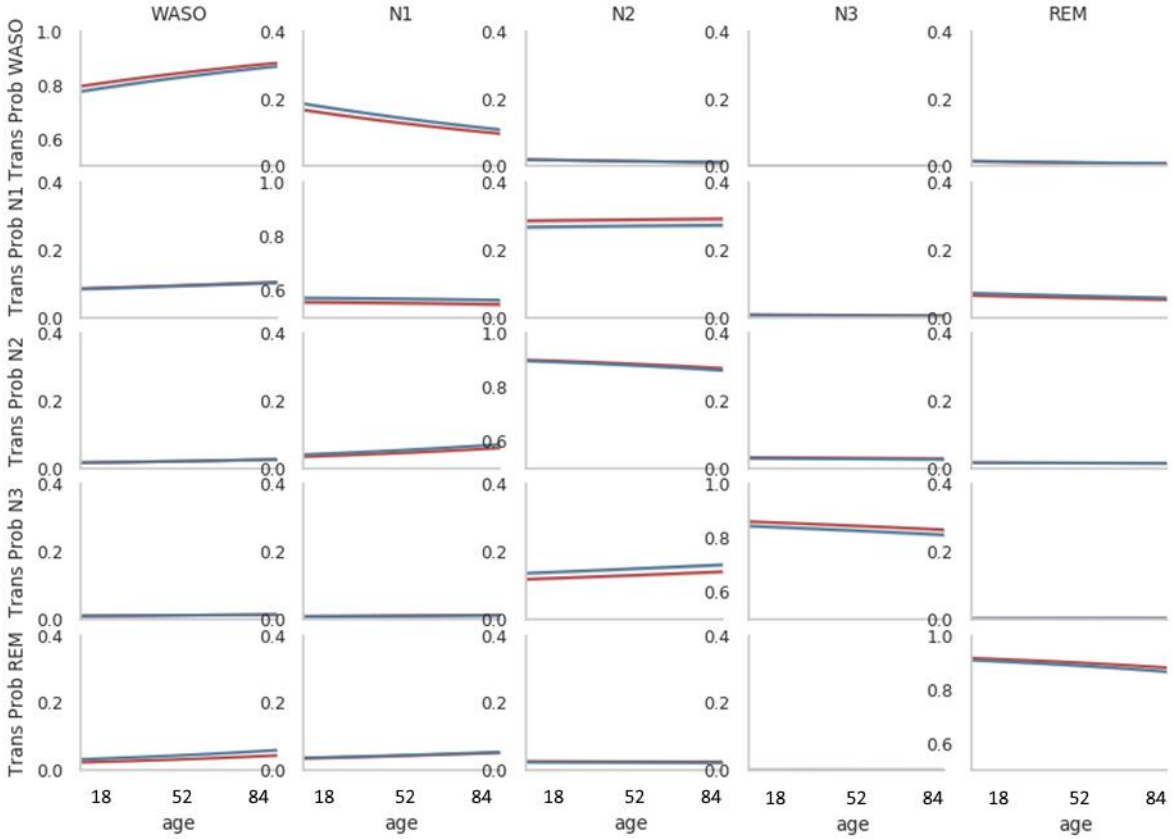

*Figure 17: Transitions probabilities from various stages (top to bottom panels) to various stages (left to right panels) as a function of age (x-axis, in years) and sex (**male in blue, female in red**). Note the different scale for self-transitions (diagonal of plot; 0.6-1) compared to non-self-transitions (off-diagonals; 0-0.4).*

Models of Sleep Features:

To determine the dynamics and influence of demographics on sleep features I fit a series of models to predict the count of each sleep feature for each epoch. With the 30 second epochs used, the magnitude of counts is half the density. Again, for computational efficiency, only a subset of the data (30,000 randomly sampled epochs) was used to determine the best model, but parameter values for the best fitting model are from the complete dataset. The Spindles and Slow Oscillation model was fit on N2 and N3 epochs in a single model (with a separate set of parameters for each feature, and a parameter for the correlation between features). The separate REM event count model was fit on REM sleep only.

Spindles and Slow Oscillations Counts

A comparison of the Spindle and SO models fit are shown in Table 4. Only the best fitting functional relationship is shown (for example, linear effects of age were tried, but quadratic was the best and is shown here).

Table 5: Model fit for models of slow oscillations and spindles

Base rate	Correl	Prev Bout	Tau	Time of Day	Age	Sex	Age X Sex	LOO	ΔLOO	Weight
Fixed	✗	✗	✗	✗	✗	✗	✗	183923		≈0
Fixed	✓	✗	✗	✗	✗	✗	✗	183922	-1	≈0
Nested	✓	✗	✗	✗	✗	✗	✗	181012	-2910	≈0
Nested	✗	✗	✗	✗	✗	✗	✗	181011	-1	≈0
Nested	✗	✓	✗	✗	✗	✗	✗	147133	-33878	≈0
Nested	✗	✓	Quadratic	✗	✗	✗	✗	143716	-3417	≈0
Nested	✗	✓	Quadratic	Quadratic	✗	✗	✗	143045	-671	≈0
Nested	✗	✓	Quadratic	Quadratic	Quadratic	✗	✗	142465	-580	<0.01
Nested	✗	✓	Quadratic	Quadratic	Quadratic	✓	✗	141286	-1179	0.05
Nested	✗	✓	Quadratic	Quadratic	Quadratic	✓	✓	141264	-22	0.94

Only the best fitting functional form (i.e. quadratic) is shown, even though all were tested. A green tick indicates the model included those variables. The red crosses indicate it does not. LOO is the leave one out cross-validation score, while ΔLOO is the difference in LOO between consecutive models. The weight column represents an estimate of the probability that the corresponding model is the best fitting model.

Interesting the best fitting model of SO and spindles did not include a correlation term between these features (however, LOO scores were very close) but did include a random intercepts term to account for dataset differences. The set of variables that best predicted these features were previous bout, a quadratic effect of tau, time of day, age and an effect of sex, and age X sex interaction term. This model was approximately 24 times better than the next best model (without age X sex interaction).

I next tested the effects of each predictor term on the expected mean of spindles and SO, while holding the other predictors at their mean. Figure 18 shows the change in spindles and SO across time in stage (tau) for N2 and N3 separately. For spindles, both N2 and N3 began with an average of 0.8 spindles per epoch (a spindle density of 1.6 spindles/min), with N2 spindles rising to 1.3 counts/epoch as the stage goes on, and N3 falling to 0.3 counts/epoch and then stabilizing. SOs showed a curious trend,

where their counts increased across the time in N2 and N3 and reached a maximum right before the transition to another stage. This rise across time is extremely pronounced for N3, with a final average SO count of approximately 9 SO per epoch after 35 minutes of continuous N3 sleep.

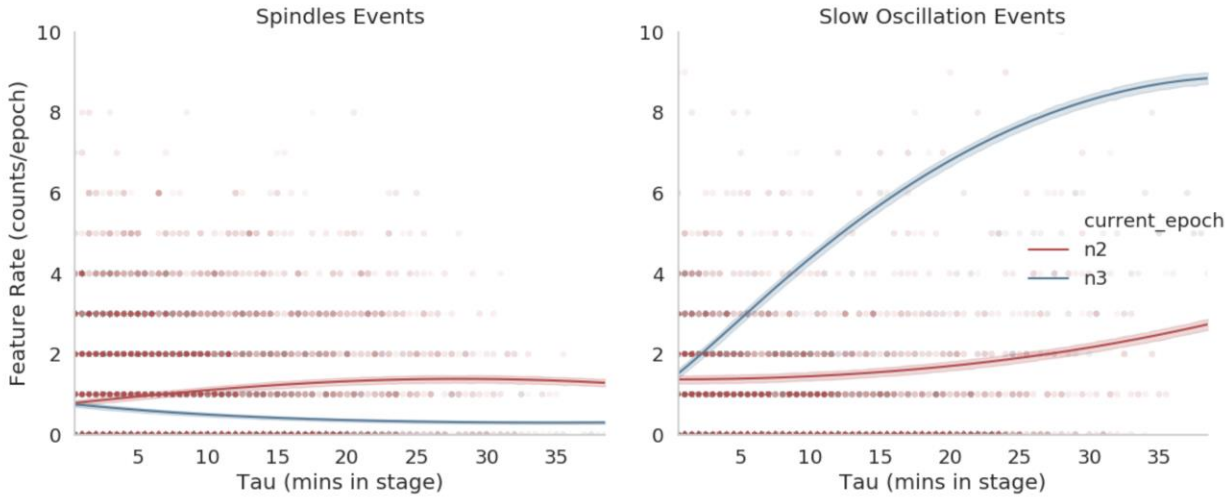

Figure 18: Counts per epoch of Spindles (left) and Slow Oscillations (right) as a function of minutes in stage (Tau) and the stage of the current epoch (n2, n3). 1000 data points with low opacity are also plotted such that darker points represent many points with the same value. Note that N2 points are plotted last, therefore final points appear red redder.

Effect of the previous bout

The previous bout was an important predictor of SO and spindle counts, where a model including the previous bout fit better than one without. Table 5 shows the percentage change between the spindle and SO counts when the previous bout is explicitly included, vs one where it is marginalized out. Small changes due to the previous stage (less than 10% change) are not included. I found N2 slow oscillations were affected by the previous bout, where N2 bouts were richer in Slow Oscillations when following wake and N3, but reduced when following REM and N1. Spindle counts in N2 increased when N2 followed REM. N3 spindle counts were reduced when N3 followed N1 and increased when N3 followed N2 (note that N2→N3 is the majority transition).

Table 6: The effect of the previous bout on feature rates

Feature	Current Epoch	Prev Bout	Feature Rate marginalizing over prev bout	Feature Rate inc. prev bout	Percent Change
Slow Osc	N2	N1	2.7	2.0	-29.0
Slow Osc	N2	N3	2.7	3.3	20.1
Slow Osc	N2	REM	2.7	1.7	-39.2
Slow Osc	N2	WASO	2.7	3.2	16.4
Spindles	N2	REM	1.1	0.8	-28.5
Spindles	N3	N1	0.5	0.3	-46.9
Spindles	N3	N2	0.5	0.7	33.3

Note that only a large change (>10%) are shown.

Effect of time of day

I then turned to look at the spindle and SO counts across the night, a relationship that was the best fit with a quadratic (Figure 19). N2 spindles remained relatively constant across the night, peaking slightly around 4 am, N3 spindles began at 0.5 spindles per epoch and rose gradually towards 1 spindle per epoch by morning. Both N2 and N3 Slow oscillations were at their peak right after sleep onset and fell towards morning. N2 dropped linearly at a rate of -0.125/hour. For N3 slow oscillations, reduction with clock time was more pronounced in the first half of the night, and SO density stabilized at 3.5 SO/epoch at around 5.30 am, where it remained till morning.

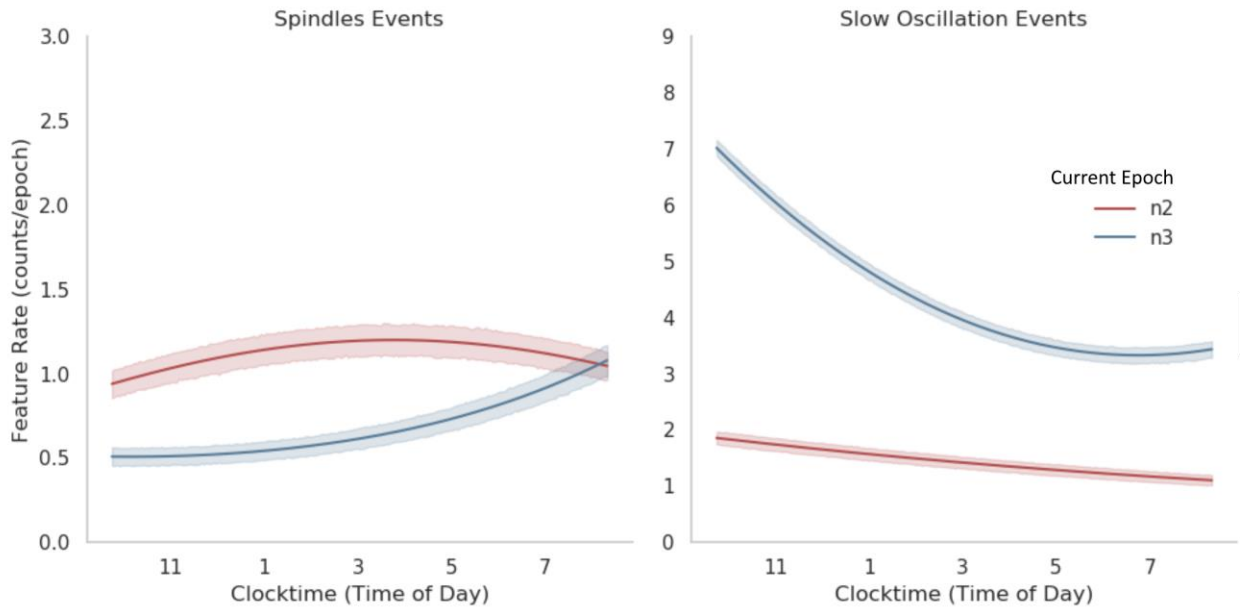

Figure 19: Counts per epoch of Spindles (left) and Slow Oscillations (right) as a function of time of day. Values above 10 represent PM, and values below AM. 1000 data points with low opacity are also plotted such that darker points

represent many points with the same value. Note that N2 points are plotted last, therefore final points appear red redder.

When I investigated the effects of age and sex in N2 (Figure 20), I saw a strong sex effect, where females have greater Spindle and SO counts than males. Spindle and SO counts both fell with age, reducing approximately 50% between 18 and 40 years. No interaction effects were present (although interactions may exist across the full age range).

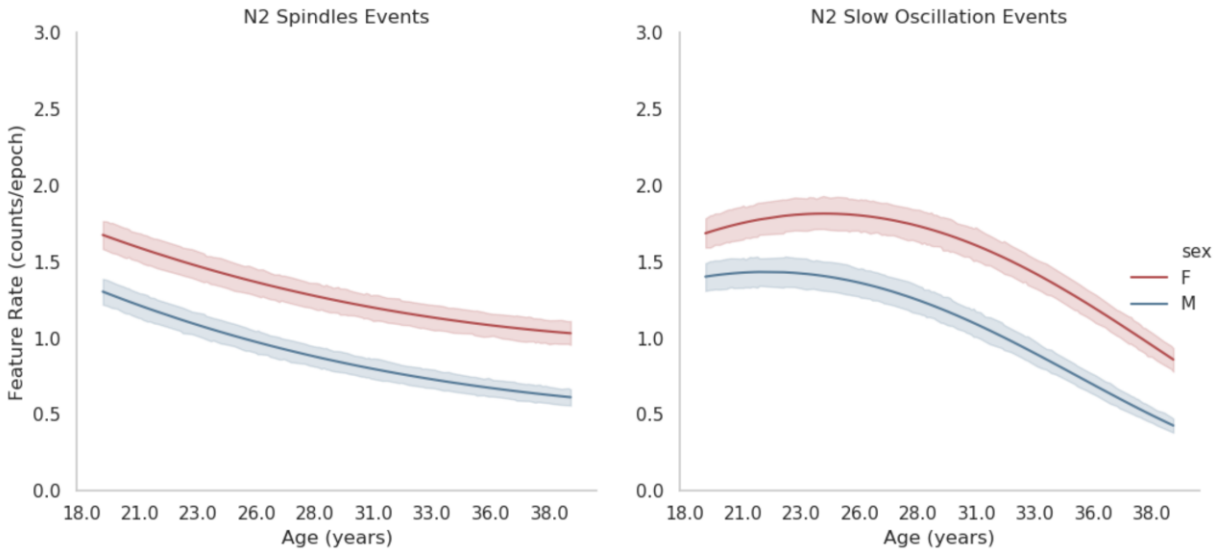

Figure 20: Counts per epoch of Spindles (left) and Slow Oscillations (right) in N2 sleep as a function of age for males (red) and females (blue) separately.

In N3, trends were quite different (Figure 21). The main effects of sex and age were lessened, but age X sex effects are seen. For N3 Spindles, females began with a higher spindle count, but this was overtaken by males towards middle age. For N3 Slow Oscillations, females began with a higher count (6 vs 4 counts/epoch) but the sexes converged as they approach middle age.

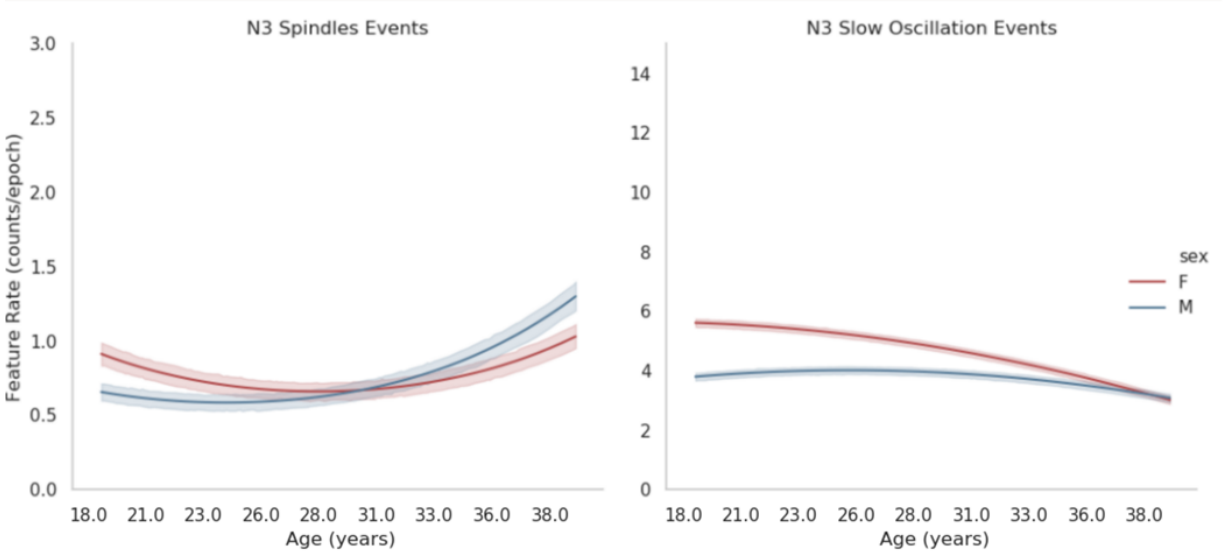

Figure 21: Counts per epoch of Spindles (left) and Slow Oscillations (right) in N3 sleep as a function of age for males (red) and females (blue) separately. Other variables held at their mean (such as time in stage).

REM Event counts

REM Density was predicted during epochs of REM only. The best-fitting model of rem density, approximately 98 times better than the next best model, consisted of a random intercepts term and accounted for REM density changes based on the previous bout, tau, time of day, age and sex (and age X sex). The model comparison table is shown in Table 6.

Table 7: Model fit metrics for the set of REM models tested

Rank	Base rate	Prev Bout	Tau	Time Of Day	Age	Sex	Age X Sex	LOO	ΔLOO	Weight
9	Fixed	✗	✗	✗	✗	✗	✗	125752		≈0
8	Random	✗	✗	✗	✗	✗	✗	125133	-619	≈0
7	Random	✓	✗	✗	✗	✗	✗	124991	-142	≈0
6	Random	✓	Quadratic	✗	✗	✗	✗	124665	-326	≈0
5	Random	✓	Quadratic	Sine	✗	✗	✗	124337	-328	≈0
4	Random	✓	Quadratic	Sine	Quadratic	✗	✗	124337	-59	≈0
3	Random	✓	Quadratic	Sine	Quadratic	✓	✗	124214	-64	0.001
2	Random	✓	Quadratic	Sine	Quadratic	✗	✗	124211	-3	0.009
1	Random	✓	Quadratic	Sine	Quadratic	✓	✓	124161	-50	0.98

Only the best fitting functional form (i.e. quadratic) is shown, even though all were tested. A green tick indicates the model included those variables. A red cross indicates it does not. LOO is the leave one out cross-validation score, while ΔLOO is the difference in LOO between consecutive models. The weight column represents an estimate of the probability that the corresponding model is the best fitting model.

From Figure 22, the mean REM density (counts/min) across all stages was 4.6, but this varied a little depending on what stage occurred prior to the current REM epoch: when the transition to REM was from WASO, or WBSO, there was a 1 and 0.58 reduction in density respectively. The transition from N1 and N2 were very close to the mean (0.08 increase, 0.04 decrease). Transitions from N3 are extremely rare, and therefore not included.

The effect of time in stage was quadratic, with up to 0.6 increase in REM counts/epoch during the middle of a REM bout.

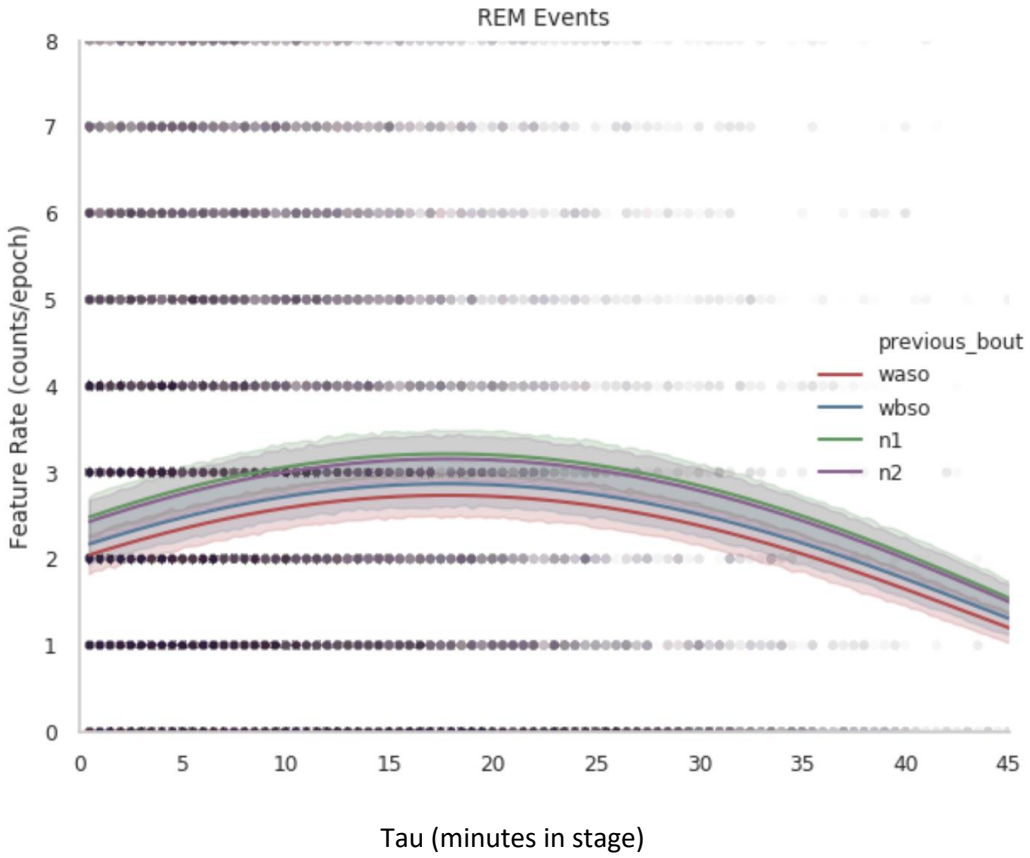

Figure 22: Counts per epoch of REM events across time in REM sleep. Wither the previous stage (before REM) was WASO, WBSO, N1 or N2 are plotted separately. N3 →REM transitions are effectively zero, and REM counts for this transition were not plotted. 1000 data points are plotted with a low opacity, and darker points represent many points with the same value.

REM density varied sinusoidally with time of day (Figure 23), with REM density constant at the start of the night, then rising gently towards morning.

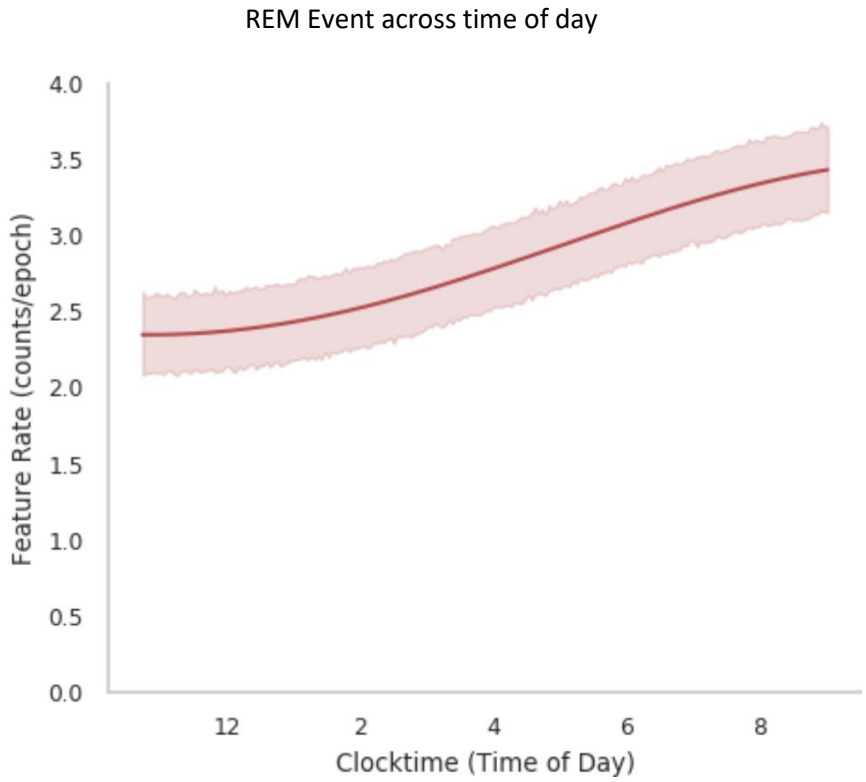

Figure 23: Counts per epoch of REM events across time of day.

Both age and sex had main effects and an interaction (Figure 24). Men's REM density begins lower but increases more with age (a linear density increase of 0.01 per year for females, and 0.03 per year for males). A positive quadratic relationship with age was observed for both sexes such that the rem density reduces slightly from young to middle age, and then begins rising throughout older age.

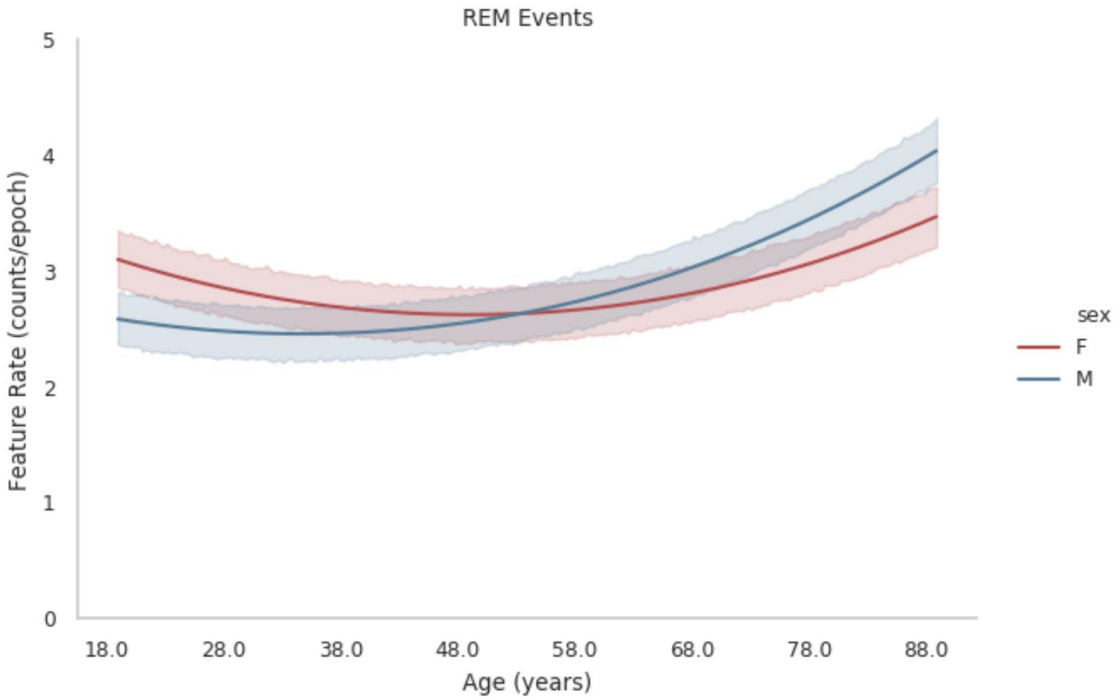

Figure 24: Counts per epoch of REM events across age for males (blue) and females (red) during REM sleep.

Discussion

Using a large multi-source dataset consisting of 1000 overnight recordings from a broad group of healthy people across a wide age range, I investigated the temporal dynamics of sleep architecture and sleep features. Modeled outcomes were the duration and pattern of sleep stages (quantified by transition probabilities) as well as the counts of spindles, slow oscillations (SO) and REM events, at an epoch-by-epoch interval. The included predictor variables were the duration of sleep stages, time across the night, and the effect of previous stages as well as the demographic variables of age and sex. Bayesian models were used to quantify all relationships, where the variance across data-sources was accounted for with a hierarchical design, and parameter estimates from prior literature were included. I tested if each predictor variable modulated the outcome variable by starting with the simplest model with only a single base rate parameter, and including variables only if they improved model fit, as measured by the model's ability to replicate out-of-sample data. By sampling the best fitting model's

parameters, and plotting outcome as a function of each input variable, I further elucidated *how* each variable influenced sleep architecture or sleep features.

I began by understanding how the pattern of sleep 5 stages – N1, N2, N3, REM, and WASO – unfolded across the night. I found, in line with previous literature and modeling work, that the pattern of sleep was affected by time of day, time in stage, age, and sex (but no age x sex interaction). Matching the work of Kishi, Bianchi, and others (Bianchi et al., 2012; Kishi et al., 2018), I found N3 and WASO stage durations were distributed paretally, whereas REM, N1, and N2 had more exponential distributions. As the night progressed, durations (i.e. self-transition probability) of WASO and N3 reduced, while all others grew. The expected rates of transitions between stages also changed from evening to morning, and transition probabilities were further modulated by the previous bout. Several interesting transition probabilities patterns were observed, for example, N2 is most likely to transition to N1, but this transition to N1 sleep is likely to be brief, where a sharp uptick in transition probability from N1 back to N2 results in numerous bouts of N2 sleep broken by very short N1 bouts – a pattern which holds across the night. I also modeled transitions from wake before sleep onset (WBSO), and found, in keeping with the literature, that the first N1 bout is expected to be longer than those in the rest of the night.

Age modulated transition probabilities, but generally to a lesser degree than time across the night, or time in stage. Durations of each stage (self-transition probabilities) were the most affected by age, where WASO durations increased across the lifespan, but all sleep stage durations fell. This points to a general fragmentation of sleep with age, where all sleep bouts are more broken, particularly by wake. Given that Conte et al found that cognitive training reduced fragmentation (Conte, Carobbi, Errico, & Ficca, 2012), an interesting extension of this model would be to include cognitive fitness as a variable, where it is predicted lead to a rise in self-transition probabilities and hence greater sleep continuity. Existing literature provides several potential explanations for increased WASO with older age, including increased need to urinate during the night (nocturia) (Coyne et al., 2003), increased

anxiety, discomfort or pain from chronic illnesses and change in hormones (such as melatonin) (Moffitt, Kalucy, Kalucy, Baum, & Cooke, 1991). In particular, previous reports indicating reduced melatonin levels (Birkeland, 1982; Sack, Lewy, Erb, Vollmer, & Singer, 1986) (a neurohormone related to circadian regulation and initiation of sleep), which decreases with age, maybe one mechanism driving older adults increased tendency to transition to WASO. Increased presence of b-amyloid and neurofibrillary tangles in the aging brain (Mander, Winer, Jagust, & Walker, 2016), which has recently be linked to sleep fragmentation (Minakawa et al., 2017), as well as cell and receptor loss in of brain areas responsible for maintaining sleep homeostasis (Lim et al., 2014; Mander et al., 2017), may also account for the previously observed disrupted sleep with age.

Sex effects on sleep stage dynamics were also detected. Women experience longer bouts of WASO, N3 and REM, similar N2 durations, and shorter N1 bouts, matching that of a previous large population-based study (Redline et al., 2004). It is intriguing that a model with an age*sex interaction term did not fit better than one without, especially considering the previous P(Sleep) 1.0 model did detect interaction effects. For example, in the previous model, Bayesian network analysis showed older males had a far higher probability to transition to WASO and a lower likelihood of SWS transition than that expected by age and sex alone. The lack of sex*age integration here maybe because of the different measures used (transition probabilities instead of stage proportions) or the more restrictive inclusion criteria employed here, such as more conservative apnea measures cutoffs. Given the propensity for apnea in older males (Gabbay & Lavie, 2012), this is a promising explanation for differences in the two modeling approaches.

Spindles, SO and REM events were affected by the full set of variables tested, including an age*sex interaction. N2 spindles and N2 slow oscillations were found to increase moderately with time spent in a stage, whereas N3 spindles and REM decreased. Slow oscillations increased rapidly across N3 sleep bouts. This increase is interesting, in that subjects with naturally shorter or fragmented N3 sleep

(but the same total N3 minutes) will have a comparatively large decrement in the number of SO across sleep. Given SO's implication in sleep-dependent memory processing, then the mechanism for reduced memory consolidation with age may be through shorter N3 bouts (found here), leading to significantly less SO's, and hence reduced memory consolidation. In keeping with this theory, fragmented sleep leads to reduced memory performance, and fragmentation increases with age (Huang et al., 2002). A study directly testing how memory performance changes as a function of SO density, N3 sleep fragmentation, and age is of high importance for future work. When considering time of day effects, Spindles (in N3) and SO were more present in the evening compared to the morning; in contrast, REM events increased over the night. These trends follow that of the stages themselves, with more NREM in the evening, and more REM sleep during the morning. However, feature counts are considered irrespective of time in stages, therefore there is both more minutes of N3 in the evening, and these minutes are richer in spindles and slow oscillations (and vice versa for REM events).

There were strong individual differences effects on sleep features. Both REM counts and N3 spindles increased with age, while N2 spindles and SO reduced. Females had more N2 spindles, and SO in both N2 and N3 than men. With less fractured SWS in women (higher self-transition probabilities) and increasing SO with time in stage, the larger number of SO for females was to be expected. However, females generally have thinner scalp thickness (Armitage & Hoffmann, 2001), and while it is tempting to conclude that they experience more SO, this may be a function of the SO detectors failing to trigger with the reduced amplitude of oscillations in men due to increased skull thickness. The patterns of N3 spindles were generally inconsistent with other NREM events (opposite effects of age, age X sex interaction, time in stage, and minutes in stage). This is evidence towards N3 spindles being functionally different from other NREM events and highlights the need to consider N2 and N3 spindles separately when analyzing their relation to memory consolidation. The effect of independent variables on REM

event patterns was similar to N3 spindles, and an investigation into the dynamic coupling between these events may be fruitful.

Sleep features present in NREM sleep – slow oscillations and spindles – were grouped into a single model. It was initially thought that there would be a correlation between the counts of these features, but the model with this correlation and did not fit better than the one without. However, a strong body of literature posits a coupling between a portion of spindles and slow oscillations (Muehlroth et al., 2019; Mohammad Niknazar, Krishnan, Bazhenov, & Mednick, 2015), and if these features co-occur, then a model with correlation should have fit better. One explanation is the correlation parameter between these features was modeled as constant with respect to other variables. The level of coupling appears to reduce with age (Helfrich, Mander, Jagust, Knight, & Walker, 2018) and the base rate of SO and spindles, and therefore coupling, is different across stages. A constant (normally distributed) value of correlation may not have been sufficient in capturing a truly fluctuating value, and the model, therefore, preferred the reduced complexity of no correlation. Another, simpler explanation is the correlation may not have been strong enough to detect through event counts.

REM sleep hallmark feature, the REM events, have not been thoroughly investigated in the past, and a novel contribution of the P(Sleep) 2.0 model is understanding their dynamics. While an age X sex interaction has been found before (Reynolds III et al., 1985; Roch et al., 1988), to my knowledge, this is the first report of a consistent rise in REM density in older age regardless of sex. This finding is in contrast to previous reports of reduced REM with age (Darchia et al., 2003), or no effect of age (Ficca et al., 1999; Peters et al., 2014; Schwarz et al., 2017). These studies did not consider a curvilinear effect of age, segment analysis by sex, or control for the factors present here (time of day, time in stage, etc). Additionally, previous studies counted REMs by hand, and the different detection procedures or statistical analysis may have contributed to the contrasting findings. It is difficult to hypothesize why REM density may increase over age, however, it may be due to a decreased sleep need (Lucidi et al.,

1996), or a rise in sub-clinical-threshold depression (Gillin et al., 1981) across the lifespan. A reciprocal relation with the amount of REM and SO/Spindles across age was observed, and a model explicitly modeling there relation within a subject would be of value. I also found REM count reduced when the previous stage was either WASO or WBSO. The lower REM density when transitioning from wake states may mean Tonic REM - a functionally quiescent state of REM sleep - is more likely after waking than Phasic REM.

Limitations

Nesting

The data used has multiple levels of a hierarchical, nested structure where subjects are nested in dataset, and there are many sleep architecture data points for a subject that share that subject's age, sex, and other unmodeled subject variables. Data points from the same subject are therefore correlated, where, for example, specific subjects may have higher (or lower) sleep feature counts than average, and using the group (dataset) mean as a predictor is a poor estimate. This is especially true for spindles, where previous studies show properties of a spindle, such as density and peak frequency, are highly subject dependent, with more variance between subjects than within (even across many nights of sleep)(De Gennaro, Ferrara, Vecchio, Curcio, & Bertini, 2005; Geiger et al., 2011). While the nesting of data points within a dataset was considered, the nested structure of data points within a subject was ignored, hence my parameter estimates may be less accurate (i.e. higher variance). The Bayesian approach employed here can incorporate more than 2 levels of hierarchy, and a rich avenue of further work is to consider individual subject level parameters, especially when modeling multiple nights from each subject.

Variables tested

P(Sleep) 2.0 did not test the effect of every variable that may influence sleep, nor the full range of possible interactions. For example, ethnicity plays a role in sleep duration: relative to whites, Blacks

are likely to be shorter sleepers (however, see (Profant, Ancoli-Israel, & Dimsdale, 2002)), and Hispanics are more likely to report extremes, e.g. extremely shorter or longer sleep duration than whites (Hale & Do, 2007; Krueger & Friedman, 2009). Other studies have noted more REM, light sleep (stage 1 and 2) and less SWS in blacks compared to whites, which may be driven by stress-induced by racial discrimination (Profant et al., 2002; Redline et al., 2004; Tomfohr, Pung, Edwards, & Dimsdale, 2012). Most subjects in this study were White, and the specific effects of ethnicity on sleep dynamics (and sleep features) remain unexplored. Future work should focus on ethnicity as a sleep moderator, especially when promoting healthy sleep through public policy. I also did not have detailed information on previous sleep history and habits, and napping, chronotypes, and circadian alignment were not considered. Napping is likely to be more apparent in the older population and may have influenced my results. In future work, the sleep and wake patterns for a single individual across many days should be considered, this data coupled with hierarchical models at the subject level would be better able to quantify the effects of daytime sleep and circadian alignment. Due to time and resource limitations, there were also interactions that were not considered, especially the time of day X individual differences. The literature investigating the Time of Day X Sex and Time of Day X Age interaction is mixed, and many have found no effect (Huang et al., 2002; Monk, Buysse, Reynolds III, Kupfer, & Houck, 1995). However, others note some changes in sleep features. For example, the intrinsic frequency of the circadian clock and the amplitude of the melatonin rhythm have been shown to differ between men and women, which in turn modulates time of day effects on Slow Wave Activity (and likely Slow Oscillations) (Santhi et al., 2016). However, this same study did not show Time of day X Sex effects on REM, NREM and wake minutes. Not all sleep features are modulated: Peters et al investigated the effect of spindles and REM across the night and noted a time of day by age interaction on REM densities but not spindles (Peters et al., 2014). By not including these interactions, some of the influence of this interaction will be

included in the main effects of sex and age, and the time of day parameters may be less accurate for each sex individually.

Distributions of outcomes

A Poisson distribution measures the rate of independent events in some continuous timeframe. The key assumption for a process to be Poisson is the independence of events in time. For this modeling work, assuming spindles and SO were Poisson distributed was a pragmatic choice that simplified modeling, however, spindles have a preferential inter-event-interval, and therefore exhibit non-Poisson-like behavior (Antony et al., 2018; Helfrich et al., 2018). Both SO and REM events also exhibit "burstiness" (Staresina et al., 2015; Yetton et al., 2016). N3 sleep can alternate between periods of high and low slow oscillations (which under older scoring systems such as R&K, would be treated as different stages), and REM alternates between Phasic and Tonic state, with high and low REM events respectively. Without modeling events in continuous time (instead of 30-second epochs), quantifying this "burstiness" of sleep features is difficult. However, a model with greater time resolution which takes refractory period into account would be of value, especially considering the refractory periods of spindles have been linked to sleep-dependent memory consolidation (Antony et al., 2018) and different cortical networks have been implicated in Phasic vs Tonic REM (Scheffzu, 2012). Along with changing the distribution of outcomes, there are additional outcomes that could be considered, such as properties of the sleep features themselves (i.e. amplitude, duration, etc.). This information is readily available from the feature detectors, and represents a natural extension of the P(Sleep) 2.0 model.

Confounds

Potential confounds exist in the data from this study (and will potentially impact the proposed study). While I excluded anyone taking sleeping medication and anti-depressants, as well as anyone with reported neurological, psychiatric or sleep disorders, other medications or illnesses may affect sleep (e.g. Beta Blockers (Betts & Alford, 1985), diabetes (Resnick et al., 2003)) and subjective sleep

complaints were not considered. It is likely that undiagnosed or subthreshold, mental, physical or sleep disorders are present in the data. These potentially unhealthy individuals likely suffer from a range of disorders with differing effects on sleep architecture, making it impossible to detect or control. Furthermore, it is likely older adults experience more of these subthreshold health issues, and the results from older subjects may be partly biased. The scoring methods used also creates a large potential confound. Sleep scored with R&K vs AASM standards have shown different sleep architecture profiles (Novelli, Ferri, & Bruni, 2010). Sleep stage models used a random intercepts model where the base-rate of sleep features (and first-order transitions could vary across dataset, which mitigated some of the additional variance from the scoring method, however, explicitly including a variable for R&K, Modified R&K, and AASM scoring methods is recommended for future work. Another issue that may impact differences across age groups is that the EEG features used to visually score sleep change with age (Landolt, Dijk, Achermann, & Borbély, 1996). Potentially, some of the observed differences in sleep in older and young adults may be due to differences in classification of sleep epochs in these populations, rather than real physiological differences.

Feature Detectors

The algorithms that were chosen here strike a balance between computational efficiency and performance. It should be noted that these algorithms are not perfect, and false-positive and negatives will have partially biased results. Both the spindle detector and REM detector used here have been previously validated by myself (Yetton et al., 2016; Yetton et al., n.d.), and the direction of this bias can be quantified. If we consider human scoring as the gold standard, then both the spindle detector and REM detector employed have a higher false-negative rate than false positive rate, and counts are likely more conservative than if human raters had detected each feature. The SO detector has not been validated rigorously, and its exact bias is unknown. All feature detectors used here were validated on a college population. A host of research has shown morphological differences in spindles and SO's with

age (Carrier et al., 2011; Yetton et al., n.d.). Many of these properties such as spindle frequency, amplitude, and duration, along with SO amplitude, are critical for feature detection (Massimini, 2004; Wamsley et al., 2012). Therefore, it was reasoned that any age-related changes in feature counts would be partly driven algorithmic miss-detection, rather than a true decline in feature counts. Spindle and SO analysis was limited to subjects 40 years and below, and the results here do not generalize out of this population. However, spindle detectors' performance may not be as poor in older adults as previous thought. My recent work in the validation of spindle algorithms found a sizeable correlation (r=0.43) of spindle density between the Wamsley algorithm (used here) and several human experts in an older population (mean age: 62 years), and a small drop in F1 score from 0.62 to 0.57. Therefore, in future work, a full age range should be considered, with the caveat that a small bias (towards fewer spindles) is expected, additionally, considering personalized thresholds for each individual, such as choosing the dominate spindles frequency from pediogram peaks (Cox, Schapiro, Manoach, & Stickgold, 2017), could reduce miss detection with age.

Conclusion and Future Directions

The body of work here introduced several novel methods in the analysis of sleep. In preliminary work, I quantified the expected patterns of sleep using Bayesian networks and a large dataset. Understanding sleep at the microscale (sleep features) is also important, and therefore I described a powerful machine learning algorithm for REM event detection. Finally, a holistic model of sleep features and sleep architecture was introduced. This model captures many variables that influence sleep patterns and acts as a high validity reference for future research questions surrounding the expected temporal patterns of healthy sleep.

This work focused on overnight sleep, both because data was readily available and its strong prior literature. However, sleep during the day, i.e. napping, is a beneficial and widely practiced pastime (Mcdevitt et al., 2018; Mednick et al., 2003). The architecture of a nap is different from that of overnight sleep, but it is unknown if this is simply driven by circadian/sleep homeostasis differences, or there are other factors at play. Including data from naps in the current modeling paradigm, while still controlling for time of day (and including an estimate of sleep pressure), would help answer this question. McDevitt et al. also found that some people consistently report feeling rested from a nap, while others eschew it due to reports of post-nap grogginess. Modeling naps at the epoch by epoch timescale may help classify the properties of restful vs non-restful naps (Mcdevitt et al., 2018).

The current work has broader impacts on the detection of abnormal (and potentially pathologic) sleep patterns. The P(Sleep) 2.0 model quantifies the likelihood of observing sleep data points under the distribution of healthy subjects. If instead, the model was fit on data-points from an unhealthy population, for example, people with depression, then the updated parameter distributions would quantify the likelihood of observing specific data-points under the distribution of depressed sleep. For a particular subject, the ratio of observing their data under the healthy model vs a depressed model could

be determined. This value, combined with the base rate of depression in the general population, could form the bases of a test for depression. The above method could be extended to a range of illnesses that have sleep effects (all most all mental illnesses, and many physical illnesses), and it an encouraging route of future research.

Finally, it should be noted that many of the finding of P(Sleep) 2.0 are not novel, but replications of numerous smaller studies. The advantage here is high confidence in conclusions due to the large sample size, stringent health criteria, and multiple variables considered. A push for larger sample sizes in recent years has been spurred by a replicability ”crisis” in science (Open Science Collaboration, 2015), and it is promising to see “big sleep data” efforts championed by the NSRR (Dean et al., 2016) and Physionet (Goldberger et al., 2000) and myself (Yetton et al., 2018; Yetton et al., n.d.). P(Sleep) 2.0 continues to move sleep towards best practices and represents a high validity summary of healthy human sleep. It is hoped that the future can leverage the methods introduced here to also help understand the patterns of unhealthy sleep, and therefore provide an important, but currently untapped, diagnosis signal.

References

Achermann, P., & Borbély, A. (1999). Sleep Homeostasis and Models of Sleep Regulation. *Wiener Medizinische Wochenschrift (1946)*, *145*, 402–406. https://doi.org/10.1111/j.1365-2869.1992.tb00013.x

Achermann, P., Dijk, D. J., Brunner, D. P., & Borbély, A. A. (1993). A model of human sleep homeostasis based on EEG slow-wave activity: quantitative comparison of data and simulations. *Brain Research Bulletin*, *31*, 97–113. https://doi.org/10.1016/0361-9230(93)90016-5

Aeschbach, D., & Borbély, A. A. (1993). All-night dynamic of the human sleep EEG. *Journal of Sleep Research*. https://doi.org/10.1111/j.1365-2869.1993.tb00065.x

Agarwal, R., Takeuchi, T., Laroche, S., & Gotman, J. (2005). Detection of rapid-eye movements in sleep studies. *IEEE Transactions on Bio-Medical Engineering*, *52*(8), 1390–1396. https://doi.org/10.1109/TBME.2005.851512

Antony, J. W., Piloto, L., Wang, M., Pacheco, P., Norman, K. A., & Paller, K. A. (2018). Sleep Spindle Refractoriness Segregates Periods of Memory Reactivation. *Current Biology*, *28*(11), 1736-1743.e4. https://doi.org/10.1016/j.cub.2018.04.020

Armitage, R., & Hoffmann, R. F. (2001). Sleep EEG, depression and gender. *Sleep Medicine Reviews*, *5*(3), 237–246. https://doi.org/10.1053/smrv.2000.0144

Aserinsky, E., & Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep, *118*(3062), 273–274. https://doi.org/10.1126/science.118.3062.273

Baker, F. C., Willoughby, A. R., Sassoon, S. A., Colrain, I. M., & de Zambotti, M. (2015). Insomnia in women approaching menopause: beyond perception. *Psychoneuroendocrinology*, *60*, 96–104. https://doi.org/10.1016/j.psyneuen.2015.06.005

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C. L., & Vaughn, B. V. (2012). The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*.

Betancourt, M., Girolami, M., & Carlo, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models. *ArXiv*.

Betts, T. A., & Alford, C. (1985). β-Blockers and sleep: a controlled trial. *European Journal of Clinical Pharmacology*, *28*, 65–68. https://doi.org/10.1007/bf00543712

Bianchi, M. T., Cash, S. S., Mietus, J., Peng, C.-K., & Thomas, R. (2010). Obstructive sleep apnea alters sleep stage transition dynamics. *PloS One*, *5*(6), e11356. https://doi.org/10.1371/journal.pone.0011356

Bianchi, M. T., Eiseman, N. A., Cash, S. S., Mietus, J., Peng, C. K., & Thomas, R. J. (2012). Probabilistic sleep architecture models in patients with and without sleep apnea. *Journal of Sleep Research*, *21*(3), 330–341. https://doi.org/10.1111/j.1365-2869.2011.00937.x

Birkeland, A. J. (1982). Plasma Melatonin Levels and Nocturnal Transitions between Sleep and Wakefulness. *Neuroendocrinology*, (34), 126–131. https://doi.org/10.1159/000123289

Bizzotto, R., Zamuner, S., De Nicolao, G., Karlsson, M. O., & Gomeni, R. (2010). Multinomial logistic estimation of Markov-chain models for modeling sleep architecture in primary insomnia patients. *Journal of Pharmacokinetics and Pharmacodynamics*, *37*(2), 137–155. https://doi.org/10.1007/s10928-009-9148-2

Blackwell, T., Yaffe, K., Ancoli-Israel, S., Redline, S., Ensrud, K. E., Stefanick, M. L., … Stone, K. L. (2011). Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study. *Journal of the American Geriatrics Society*, *59*(12), 2217–2225. https://doi.org/10.1111/j.1532-5415.2011.03731.x

Blank, J. B., Cawthon, P. M., Carrion-Petersen, M., Harper, L., Johnson, J. P., Mitson, E., & Delay, R. R. (2005). Overview of recruitment for the osteoporotic fractures in men study (MrOS). *Contemporary Clinical Trials*, *26*(5), 557–568.

Boostani, R., Sadatnezhad, K., & Sabeti, M. (2009). An efficient classifier to diagnose of schizophrenia based on the EEG signals Expert Systems with Applications An efficient classifier to diagnose of schizophrenia based on the EEG signals, (April). https://doi.org/10.1016/j.eswa.2008.07.037

Borbély, A. A. (1982). A two process model of sleep regulation. *Human Neurobiology*.

Bove, T., Culebras, H., Moore, T., & Westlake, E. (1994). Clinical Research Relationship Between Sleep Spindles and Hypersomnia. *Sleep*, *17*(5), 449–455.

Carrier, J., Viens, I., Poirier, G., Robillard, R., Lafortune, M., Vandewalle, G., … Filipini, D. (2011). Sleep slow wave changes during the middle years of life. *European Journal of Neuroscience*, *33*(4), 758–766.

Carskadon, M. A., & Dement, W. C. (2005). Normal human sleep: an overview. *Principles and Practice of Sleep Medicine*, *4*, 13–23.

Carvalho, D. Z., Gerhardt, G. J. L., Lemke, N., Schönwald, S. V, & de Santa Helena, E. L. (2012). Topography-specific spindle frequency changes in Obstructive Sleep Apnea. *BMC Neuroscience*, *13*(1), 89. https://doi.org/10.1186/1471-2202-13-89

Cellini, N., Goodbourn, P. T., McDevitt, E. a, Martini, P., Holcombe, A. O., & Mednick, S. C. (2015). Sleep after practice reduces the attentional blink. *Attention, Perception & Psychophysics*, *77*(6), 1945–1954. https://doi.org/10.3758/s13414-015-0912-7

Chambon, S., Thorey, V., Arnal, P. J., Mignot, E., Gramfort, A., & Neurospin, C. E. A. (2018). A deep learning architecture to detect events in EEG signals during sleep. *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Chu-Shore, J., Westover, M. B., & Bianchi, M. T. (2010). Power law versus exponential state transition dynamics: Application to sleep-wake architecture. *PLoS ONE*, *5*(12). https://doi.org/10.1371/journal.pone.0014204

Cohen-Zion, M., Drummond, S. P. A., Padula, C. B., Winward, J., Kanady, J., Medina, K. L., & Tapert, S. F. (2009). Sleep architecture in adolescent marijuana and alcohol users during acute and extended abstinence. *Addictive Behaviors*, *34*(11), 976–979. https://doi.org/10.1016/j.addbeh.2009.05.011

Conte, F., Arzilli, C., Errico, B. M., Giganti, F., Iovino, D., & Ficca, G. (2014). Sleep Measures Expressing "Functional Uncertainty" in Elderlies' Sleep. *Gerontology*, *60*(5), 448–457.

https://doi.org/10.1159/000358083

Conte, F., Carobbi, G., Errico, B. M., & Ficca, G. (2012). The effects of pre-sleep learning on sleep continuity , stability , and organization in elderly individuals, *3*(July), 1–9. https://doi.org/10.3389/fneur.2012.00109

Cooper, G. F., & Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*(4), 309–347. https://doi.org/10.1007/BF00994110

Cox, R., Schapiro, A. C., Manoach, D. S., & Stickgold, R. (2017). Individual Differences in Frequency and Topography of Slow and Fast Sleep Spindles. *Frontiers in Human Neuroscience*, *11*, 433. https://doi.org/10.3389/fnhum.2017.00433

Coyne, K. S., Zhou, Z., Bhattacharyya, S. K., Thompson, C. L., Dhawan, R., & Versi, E. (2003). The prevalence of nocturia and its effect on health-related quality of life and sleep in a community sample in the USA. *BJU International*, *92*(9), 948–954. https://doi.org/10.1111/j.1464-410X.2003.04527.x

Daan, S., Beersma, D. G., & Borbély, A. (1984). Timing of human sleep: recovery process gated by a circadian pacemaker. *The American Journal of Physiology*, *246*(2 Pt 2), R161–R183.

Darchia, N., Campbell, I. G., & Feinberg, I. (2003). Rapid eye movement density is reduced in the normal elderly. *Sleep*, *26*(8), 973–977.

De Gennaro, L., Ferrara, M., Vecchio, F., Curcio, G., & Bertini, M. (2005). An electroencephalographic fingerprint of human sleep. *NeuroImage*, *26*(1), 114–122. https://doi.org/10.1016/j.neuroimage.2005.01.020

De Oliveira Alvares, L., Crestani, A. P., Cassini, L. F., Haubrich, J., Santana, F., & Quillfeldt, J. A. (2013). Reactivation enables memory updating, precision-keeping and strengthening: Exploring the possible biological roles of reconsolidation. *Neuroscience*, *244*, 42–48. https://doi.org/10.1016/j.neuroscience.2013.04.005

Dean, D. A., Goldberger, A. L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., … others. (2016). Scaling up scientific discovery in sleep medicine: The National Sleep Research Resource. *Sleep*, *39*(5), 1151–1164. https://doi.org/10.5665/sleep.5774

Doman, J., Detka, C., Hoffman, T., Kesicki, D., Monahan, J. P., Buysse, D. J., … Kupfer, D. J. (1995). Automating the sleep laboratory: Implementation and validation of digital recording and analysis. *International Journal of Bio-Medical Computing*, *38*(3), 277–290. https://doi.org/10.1016/S0020-7101(05)80010-8

Drake, C. L., Roehrs, T., & Roth, T. (2003). Insomnia causes, consequences, and therapeutics: An overview. *Depression and Anxiety*, *18*(4), 163–176. https://doi.org/10.1002/da.10151

Drake, C., Roehrs, T., Shambroom, J., & Roth, T. (2013). Caffeine effects on sleep taken 0, 3, or 6 hours before going to bed. *Journal of Clinical Sleep Medicine*, *9*(11), 1195–1200. https://doi.org/10.5664/jcsm.3170

Dubé, J., Lafortune, M., Bedetti, C., Bouchard, M., Gagnon, J. F., Doyon, J., … Carrier, J. (2015). Cortical thinning explains changes in sleep slow waves during adulthood. *Journal of Neuroscience*, *35*(20), 7795–7807.

Ehlers, C. L., & Kupfer, D. J. (1989). Effects of age on delta and REM sleep parameters. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 118–125.

Ehlers, C. L., & Kupfer, D. J. (1997). Slow-wave sleep: do young adult men and women age differently? *J Sleep Res.*, 211–215.

El-Ad, B., & Lavie, P. (2005). Effect of sleep apnea on cognition and mood. *International Review of Psychiatry*, *17*(4), 277–282.

Ferrillo, F., Donadio, S., De Carli, F., Garbarino, S., & Nobili, L. (2007). A model-based approach to homeostatic and ultradian aspects of nocturnal sleep structure in narcolepsy. *Sleep*, *30*(2), 157–165.

Ficca, G., Gori, S., Ktonas, P., Quattrini, C., Trammell, J., & Salzarulo, P. (1999). The organization of rapid eye movement activity during rapid eye movement sleep is impaired in the elderly. *Neuroscience Letters*, *275*(3), 219–221.

Freund, Y., & Schapire, R. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*(5), 771–780. https://doi.org/citeulike-article-id:765005

Gabbay, I. E., & Lavie, P. (2012). Age-and gender-related characteristics of obstructive sleep apnea. *Sleep and Breathing*, *16*(2), 453–460.

Geiger, A., Huber, R., Kurth, S., Ringli, M., Jenni, O. G., & Achermann, P. (2011). The sleep EEG as a marker of intellectual ability in school age children. *Sleep*, *34*(2), 181–189. https://doi.org/10.5167/uzh-44595

Genzel, L., Spoormaker, V. I., Konrad, B. N., & Dresler, M. (2015). The role of rapid eye movement sleep for amygdala-related memory processing. *Neurobiology of Learning and Memory*, *122*, 101–121. https://doi.org/10.1016/j.nlm.2015.01.008

Gillin, J. C., Duncan, W. C., Murphy, D. L., Post, R. M., Wehr, T. A., Goodwin, F. K., … Bunney Jr, W. E. (1981). Age-related changes in sleep in depressed and normal subjects. *Psychiatry Research*, *4*(1), 73–78.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., … Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet. *Circulation*, *101*(23), e215--e220.

Groch, S., Wilhelm, I., Diekelmann, S., & Born, J. (2013). The role of REM sleep in the processing of emotional memories: Evidence from behavior and event-related potentials. *Neurobiology of Learning and Memory*, *99*, 1–9. https://doi.org/10.1016/j.nlm.2012.10.006

Hale, L., & Do, D. P. (2007). Racial differences in self-reports of sleep duration in a population-based study. *Sleep*, *30*(9), 1096–1103. https://doi.org/10.1093/sleep/30.9.1096

Hatzilabrou, G. M., Greenberg, N., Sclabassi, R. J., Carroll, T., Guthrie, R. D., & Scher, M. S. (1994). A comparison of conventional and matched filtering techniques for rapid eye movement detection of the newborn. *IEEE Transactions on Bio-Medical Engineering*, *41*(10), 990–995. https://doi.org/10.1109/10.324532

Helfrich, R. F., Mander, B. A., Jagust, W. J., Knight, R. T., & Walker, M. P. (2018). Old Brains Come Uncoupled in Sleep: Slow Wave-Spindle Synchrony, Brain Atrophy, and Forgetting. *Neuron*, *97*(1), 221-230.e4. https://doi.org/10.1016/j.neuron.2017.11.020

Hibbs, A. M., Storfer-Isser, A., Rosen, C., Ievers-Landis, C. E., Taveras, E. M., & Redline, S. (2014). Advanced sleep phase in adolescents born preterm. *Behavioral Sleep Medicine*, *12*(5), 412–424.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Huang, Y.-L., Liu, R.-Y., Wang, Q.-S., Van Someren, E. J. W., Xu, H., & Zhou, J.-N. (2002). Age-associated difference in circadian sleep--wake and rest--activity rhythms. *Physiology & Behavior*, *76*(4–5), 597–603.

Hutchison, I. C., & Rathore, S. (2015). The role of REM sleep theta activity in emotional memory. *Frontiers in Psychology*, *6*(OCT), 1–15. https://doi.org/10.3389/fpsyg.2015.01439

Iber, C., Ancoli-Israel, S., Chesson, A., Quan, S. F., & others. (2007). *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications* (Vol. 1). American Academy of Sleep Medicine Westchester, IL.

Karlsson, M. O., Schoemaker, R. C., Kemp, B., Cohen, A. F., Van Gerven, J. M. A., Tuk, B., … Danhof, M. (2000). A pharmacodynamic Markov mixed-effects model for the effect of temazepam on sleep. *Clinical Pharmacology and Therapeutics*, *68*(2), 175–188. https://doi.org/10.1067/mcp.2000.108669

Kemp, B., & Kamphuisen, H. A. C. (1986). Simulation of Human Hypnograms Using a Markov Chain Model. *Sleep*, *9*(January), 405–414.

Khalsa, S. B. S., Conroy, D. A., Duffy, J. F., Czeisler, C. A., & DIJK, D.-J. (2002). Sleep-and circadian-dependent modulation of REM density. *Journal of Sleep Research*, *11*(1), 53–59.

Kim, J. W., Lee, J. S., Robinson, P. A., & Jeong, D. U. (2009). Markov analysis of sleep dynamics. *Physical Review Letters*, *102*(17), 1–4. https://doi.org/10.1103/PhysRevLett.102.178104

Kishi, A., Natelson, B. H., Togo, F., Struzik, Z. R., Rapoport, D. M., & Tamamoto, Y. (2011). Sleep-stage dynamics in patients with Chronic Fatigue Syndrome with or without Fibromyalgia. *Sleep*, *34*(11), 1551–1560.

Kishi, A., Struzik, Z. R., Natelson, B. H., Togo, F., & Yamamoto, Y. (2008). Dynamics of sleep stage transitions in healthy humans and patients with chronic fatigue syndrome. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, *294*(6), R1980–R1987. https://doi.org/10.1152/ajpregu.00925.2007

Kishi, A., Yamaguchi, I., Togo, F., & Yamamoto, Y. (2018). Markov modeling of sleep stage transitions and ultradian REM sleep rhythm. *Physiological Measurement*, *39*(8). https://doi.org/10.1088/1361-6579/aad900

Kishi, A., Yasuda, H., Matsumoto, T., Inami, Y., Horiguchi, J., Tamaki, M., … Yamamoto, Y. (2011). NREM Sleep Stage Transitions Control Ultradian REM Sleep Rhythm. *Sleep*, *34*(10), 1423–1432. https://doi.org/10.5665/Sleep.1292

Kovács, E., Kosztolányi, A., & Kis, A. (2018). Rapid eye movement density during REM sleep in dogs (Canis familiaris), 554–560.

Krueger, P. M., & Friedman, E. M. (2009). Sleep duration in the united states: A cross-sectional population-based study. *American Journal of Epidemiology*, *169*(9), 1052–1063.

https://doi.org/10.1093/aje/kwp023

Ktonas, P. Y., & Smith, J. R. (1978). Automatic rem detection: Modifications on an existing system and preliminary normative data. *International Journal of Bio-Medical Computing*, *9*(6), 445–464. https://doi.org/10.1016/0020-7101(78)90052-1

Kupfer, D. J. (1981). Application of EEG sleep in the treatment of depression. In *Electroneurophysiology and Psychopathology* (pp. 87–93). Karger Publishers.

Lacourse, K., Delfrate, J., Beaudry, J., Peppard, P., & Warby, S. C. (2018). A sleep spindle detection algorithm that emulates human expert spindle scoring. *Journal of Neuroscience Methods*, (August), 1–9. https://doi.org/10.1016/j.jneumeth.2018.08.014

Lajnef, T., Chaibi, S., Eichenlaub, J., Ruby, P. M., Aguera, P., Samet, M., & Kachouri, A. (2015). Sleep spindle and K-complex detection using tunable Q-factor wavelet transform and morphological component analysis, *9*(July), 1–17. https://doi.org/10.3389/fnhum.2015.00414

Landolt, H. P., Dijk, D. J., Achermann, P., & Borbély, A. A. (1996). Effect of age on the sleep EEG: Slow-wave activity and spindle frequency activity in young and middle-aged men. *Brain Research*, *738*(2), 205–212. https://doi.org/10.1016/S0006-8993(96)00770-6

Latta, F., Leproult, R., Tasali, E., Hofmann, E., L'Hermite-Balériaux, M., Copinschi, G., & Van Cauter, E. (2005). Sex Differences in Nocturnal Growth Hormone and Prolactin Secretion in Healthy Older Adults: Relationships With Sleep EEG Variables. *Sleep*, *28*(12), 1519. https://doi.org/10.1093/sleep/28.12.1519

Lim, A. S. P., Ellison, B. A., Wang, J. L., Yu, L., Schneider, J. A., Buchman, A. S., … Saper, C. B. (2014). Sleep is related to neuron numbers in the ventrolateral preoptic/intermediate nucleus in older adults with and without Alzheimer's disease. *Brain*, *137*(10), 2847–2861.

Lo, C.-C., Chou, T., Penzel, T., Scammell, T. E., Strecker, R. E., Stanley, H. E., & Ivanov, P. C. (2004). Common scale-invariant patterns of sleep-wake transitions across mammalian species. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(50), 17545–17548. https://doi.org/10.1073/pnas.0408242101

Lo, C., Nunes Amaral, L. A., Havlin, S., Ivanov, P. C., Penzel, T., Peter, J. H., & Stanley, H. E. (2002). Dynamics of sleep-wake transitions during sleep.pdf. *Europhys. Lett.*, *57*(5), 625–631.

Lucidi, F., Devoto, A., Violani, C., De Gennaro, L., Mastracci, P., & Bertini, M. (1996). Rapid eye movements density as a measure of sleep need: REM density decreases linearly with the reduction of prior sleep duration. *Electroencephalography and Clinical Neurophysiology*, *99*(6), 556–561.

Mander, B. A., Rao, V., Lu, B., Saletin, J. M., Ancoli-Israel, S., Jagust, W. J., & Walker, M. P. (2014). Impaired prefrontal sleep spindle regulation of hippocampal-dependent learning in older adults. *Cerebral Cortex*, *24*(12), 3301–3309. https://doi.org/10.1093/cercor/bht188

Mander, B. A., Winer, J. R., Jagust, W. J., & Walker, M. P. (2016). Sleep: A Novel Mechanistic Pathway, Biomarker, and Treatment Target in the Pathology of Alzheimer's Disease? *Trends in Neurosciences*, *39*(8), 552–566. https://doi.org/10.1016/j.tins.2016.05.002

Mander, B. A., Winer, J. R., & Walker, M. P. (2017). Review Sleep and Human Aging. *Neuron*, *94*(1), 19–36. https://doi.org/10.1016/j.neuron.2017.02.004

Mander, B. a, Marks, S. M., Vogel, J. W., Rao, V., Lu, B., Saletin, J. M., … Walker, M. P. (2015). β-amyloid disrupts human NREM slow waves and related hippocampus-dependent memory consolidation. *Nature Neuroscience*, (June), 1–10. https://doi.org/10.1038/nn.4035

Manoach, D. S., Pan, J. Q., Purcell, S. M., & Stickgold, R. (2016). Reduced Sleep Spindles in Schizophrenia: A Treatable Endophenotype That Links Risk Genes to Impaired Cognition? *Biological Psychiatry*, *80*(8), 599–608. https://doi.org/10.1016/j.biopsych.2015.10.003

Martin, N., Lafortune, M., Godbout, J., Barakat, M., Robillard, R., Poirier, G., … Carrier, J. (2013). Topography of age-related changes in sleep spindles. *Neurobiology of Aging*, *34*(2), 468–476.

Massimini, M. (2004). The Sleep Slow Oscillation as a Traveling Wave. *Journal of Neuroscience*, *24*(31), 6862–6870. https://doi.org/10.1523/JNEUROSCI.1318-04.2004

McArdle, N., & Douglas, N. J. (2001). Effect of continuous positive airway pressure on sleep architecture in the sleep apnea--hypopnea syndrome: a randomized controlled trial. *American Journal of Respiratory and Critical Care Medicine*, *164*(8), 1459–1463.

Mcdevitt, E. A., Sattari, N., Duggan, K. A., Cellini, N., Whitehurst, L. N., Perera, C., … Mednick, S. C. (2018). The impact of frequent napping and nap practice on sleep-dependent memory in humans, (June), 1–12. https://doi.org/10.1038/s41598-018-33209-0

McPartland, R. J., Kupfer, D. J., & Foster, F. G. (1973). Rapid eye movement analyzer. *Electroencephalography and Clinical Neurophysiology*, *34*(3), 317–320.

Mednick, S., McDevitt, E. A., Walsh, J. K., Wamsley, E., Paulus, M., Kanady, J. C., & Drummond, S. P. A. (2013). The Critical Role of Sleep Spindles in Hippocampal-Dependent Memory: A Pharmacology Study. *J Neurosci*, *33*(10), 4494–4504. https://doi.org/10.1523/JNEUROSCI.3127-12.2013

Mednick, S., Nakayama, K., & Stickgold, R. (2003). Sleep-dependent learning: a nap is as good as a night, *6*(7), 697–698. https://doi.org/10.1038/nn1078

Merica, H., & Blois, R. (1991). Relationship between the time courses of power in the frequency bands of human sleep EEG, 116–128.

Minakawa, E. N., Miyazaki, K., Maruo, K., Yagihara, H., Fujita, H., Wada, K., & Nagai, Y. (2017). Chronic sleep fragmentation exacerbates amyloid β deposition in Alzheimer's disease model mice. *Neuroscience Letters*, *653*, 362–369. https://doi.org/10.1016/j.neulet.2017.05.054

Minard, J. G., & Krausman, D. (1971). RAPID EYE MOVEMENT DEFINITION A N D COUNT :, 99–102.

Moffitt, P. F., Kalucy, E. C., Kalucy, R. S., Baum, F. E., & Cooke, R. D. (1991). Sleep difficulties, pain and other correlates. *Journal of Internal Medicine*, *230*(3), 245–249.

Monk, T. H., Buysse, D. J., Reynolds III, C. F., Kupfer, D. J., & Houck, P. R. (1995). Circadian temperature rhythms of older people. *Experimental Gerontology*, *30*(5), 455–474.

Muehlroth, B. E., Sander, M. C., Fandakova, Y., Grandy, T. H., Rasch, B., Shing, Y. L., & Werkle-bergner, M. (2019). Precise Slow Oscillation – Spindle Coupling Promotes Memory Consolidation in Younger and Older Adults, (November 2018), 1–15. https://doi.org/10.1038/s41598-018-36557-z

Niknazar, M, Mousavi, S. R., Motaghi, S., Dehghani, a, Vosoughi Vahdat, B., Shamsollahi, M. B., … Noorbakhsh, S. M. (2013). A unified approach for detection of induced epileptic seizures in rats using ECoG signals. *Epilepsy & Behavior : E&B*, *27*(2), 355–364.

https://doi.org/10.1016/j.yebeh.2013.01.028

Niknazar, Mohammad, Krishnan, G. P., Bazhenov, M., & Mednick, S. C. (2015). Coupling of Thalamocortical Sleep Oscillations Are Important for Memory Consolidation in Humans. *PLoS ONE*, *10*(12), 1–14. https://doi.org/10.1371/journal.pone.0144720

Novelli, L., Ferri, R., & Bruni, O. (2010). Sleep classification according to AASM and Rechtschaffen and Kales: Effects on sleep scoring parameters of children and adolescents. *Journal of Sleep Research*, *19*(1 PART. 2), 238–247. https://doi.org/10.1111/j.1365-2869.2009.00785.x

O'Reilly, C., Gosselin, N., Carrier, J., & Nielsen, T. (2014). Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research. *Journal of Sleep Research*, *23*(6), 628–635. https://doi.org/10.1111/jsr.12169

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. V. (2003). Meta-Analysis of Quantitative Sleep Parameters From Childhood to Old Age in Healthy Individuals : Developing Normative Sleep Values Across the Human Lifespan. *Sleep*, *31*(2), 20–23.

Open-Science-Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Orwoll, E., Blank, J. B., Barrett-Connor, E., Cauley, J., Cummings, S., Ensrud, K., … others. (2005). Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study—a large observational study of the determinants of fracture in older men. *Contemporary Clinical Trials*, *26*(5), 569–585.

Peters, K. R., Ray, L. B., Fogel, S., Smith, V., & Smith, C. T. (2014). Age differences in the variability and distribution of sleep spindle and rapid eye movement densities. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0091047

Plante, D. T., Goldstein, M. R., Landsness, E. C., Peterson, M. J., Riedner, B. A., Ferrarelli, F., … Benca, R. M. (2013). Topographic and sex-related differences in sleep spindles in major depressive disorder: A high-density EEG investigation. *Journal of Affective Disorders*, *146*(1), 120–125. https://doi.org/10.1016/j.jad.2012.06.016

Profant, J., Ancoli-Israel, S., & Dimsdale, J. E. (2002). Are there ethnic differences in sleep architecture? *American Journal of Human Biology*, *14*(3), 321–326. https://doi.org/10.1002/ajhb.10032

Purcell, S. M., Manoach, D. S., Demanuele, C., Cade, B., Mariani, S., Cox, R., … Stickgold, R. (2016). Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource. *Preprint*, *8*(May), 1–16. https://doi.org/10.1038/ncomms15930

Quan, S., Howard, B., Iber, C., Kiley, J., Nieto, F., O'Connor, G., … Wahl, P. (1997). The Sleep Heart Health Study: design, rationale, and methods. *Sleep*, *20*(12), 1077–1085.

Rao, M. N., Blackwell, T., Redline, S., Stefanick, M. L., Ancoli-Israel, S., & Stone, K. L. (2009). Association between sleep architecture and measures of body composition. *Sleep*, *32*(4), 483–490.

Redline, S., Kirchner, H. L., Quan, S. F., Gottlieb, D. J., Kapur, V., & Newman, A. (2004). The Effects of Age, Sex, Ethnicity, and Sleep-Disordered Breathing on Sleep Architecture. *Arch Intern Med*, *164*.

Redline, S., Sanders, M. H., Lind, B. K., Quan, S. F., Iber, C., Gottlieb, D. J., … others. (1998). Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. *Sleep*, *21*(7),

759–768.

Resnick, H. E., Redline, S., Shahar, E., Gilpin, A., Newman, A., Walter, R., … Punjabi, N. M. (2003). Diabetes and sleep disturbances: Findings from the Sleep Heart Health Study. *Diabetes Care*, *26*(3), 702–709. https://doi.org/10.2337/diacare.26.3.702

Reynolds III, C. F., Kupfer, D. J., Taska, L. S., Hoch, C. C., Sewitch, D. E., & Spiker, D. G. (1985). Sleep of healthy seniors: a revisit. *Sleep*, *8*(1), 20–29.

Roch, C. C., Reynolds III, C. F., Kupfer, D. J., & Berman, S. R. (1988). Stability of EEG sleep and sleep quality in healthy seniors. *Sleep*, *11*(6), 521–527.

Sack, R. L., Lewy, A. J., Erb, D. L., Vollmer, W. M., & Singer, C. M. (1986). Human Melatonin Production Decreases With Age. *Journal of Pineal Research*, *3*(4), 379–388. https://doi.org/10.1111/j.1600-079X.1986.tb00760.x

Salvatier, J., Wiecki, T. V, & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Santhi, N., Lazar, A. S., McCabe, P. J., Lo, J. C., Groeger, J. A., & Dijk, D.-J. (2016). Sex differences in the circadian regulation of sleep and waking cognition in humans. *Proceedings of the National Academy of Sciences*, *113*(19), E2730--E2739.

Sattari, N., McDevitt, E. A., Panas, D., Niknazar, M., Ahmadi, M., Naji, M., … Mednick, S. C. (n.d.). The Effect of Sex and Menstrual Phase on Memory Formation during Nap. *Neurobiology of Learning and Memory.*

Scheffzu, C. (2012). Phasic vs tonic REM sleep in mice Distinct features of fast oscillations in phasic and tonic rapid eye movement sleep, 630–633. https://doi.org/10.1111/j.1365-2869.2012.01037.x

Schlemmer, A., Parlitz, U., Luther, S., Wessel, N., & Penzel, T. (2015). Changes of sleep-stage transitions due to ageing and sleep disorder. *Phil. Trans. R. Soc.*, *A373*(20140093), 1–16. https://doi.org/http://dx.doi.org/10.1098/rsta.2014.0093

Schwarz, J. F. A., Åkerstedt, T., Lindberg, E., Gruber, G., Fischer, H., & Theorell-Haglöw, J. (2017). Age affects sleep microstructure more than sleep macrostructure. *Journal of Sleep Research*, *26*(3), 277–287.

Shokrollahi, P., Krishnan, S., Umapathy, K., McConville, K., Boulos, M. I., Jewell, D., & Murray, B. J. (2009). Computer-assisted method for quantifying sleep eye movements that reflects medication effects. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, 1347–1350. https://doi.org/10.1109/IEMBS.2009.5334147

Spilsbury, J. C., Storfer-Isser, A., Drotar, D., Rosen, C. L., Kirchner, H. L., & Redline, S. (2005). Effects of the home environment on school-aged children's sleep. *Sleep*, *28*(11), 1419–1427.

Staresina, B. P., Bergmann, T. O., Bonnefond, M., Meij, R. Van Der, Jensen, O., Deuker, L., … Fell, J. (2015). Hierarchical nesting of slow oscillations , spindles and ripples in the human hippocampus during sleep. *Nature Publishing Group*, *18*(11), 1679–1686. https://doi.org/10.1038/nn.4119

Taheri, S., Lin, L., Austin, D., Young, T., & Mignot, E. (2004). Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. *PLoS Medicine*, *1*(3), 210–217.

https://doi.org/10.1371/journal.pmed.0010062

The MathWorks Inc. (2015). MATLAB. Natick, Massachusetts, United States.

Tomfohr, L., Pung, M. A., Edwards, K. M., & Dimsdale, J. E. (2012). Racial differences in sleep architecture: The role of ethnic discrimination. *Biological Psychology*, *89*(1), 34–38. https://doi.org/10.1016/j.biopsycho.2011.09.002

Tsuji, Y., Satoh, H., Itoh, N., Sekiguchi, Y., & Nagasawa, K. (2000). Automatic detection of rapid eye movements by discrete wavelet transform. *Psychiatry and Clinical Neurosciences*, *54*(3), 276–277. https://doi.org/10.1046/j.1440-1819.2000.00676.x

Ujma, P. P., Bódizs, R., Gombos, F., Stintzing, J., Konrad, B. N., Genzel, L., … Dresler, M. (2015). Nap sleep spindle correlates of intelligence. *Scientific Reports*, *5*, 17159. https://doi.org/10.1038/srep17159

Varri. (1996). Nonlinear Eye Movement Detector Drowsyness, *43*, 227–242.

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, (September), 1–20. https://doi.org/10.1007/s11222-016-9696-4

Wamsley, E. J., Hamilton, K., Graveline, Y., Manceor, S., & Parr, E. (2016). Test Expectation Enhances Memory Consolidation across Both Sleep and Wake. *Plos One*, *11*(10), e0165141. https://doi.org/10.1371/journal.pone.0165141

Wamsley, E. J., Tucker, M. A., Shinn, A. K., Ono, K. E., McKinley, S. K., Ely, A. V., … Manoach, D. S. (2012). Reduced sleep spindles and spindle coherence in schizophrenia: Mechanisms of impaired memory consolidation? *Biological Psychiatry*, *71*(2), 154–161. https://doi.org/10.1016/j.biopsych.2011.08.008

Warby, S. C., Wendt, S. L., Welinder, P., Munk, E. G. S., Carrillo, O., Sorensen, H. B. D., … Mignot, E. (2014). Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods*, *11*(4), 385–392. https://doi.org/10.1038/nmeth.2855

Wei, Y., Colombo, M. A., Ramautar, J. R., Blanken, T. F., van der Werf, Y. D., Spiegelhalder, K., … Van Someren, E. J. W. (2017). Sleep stage transition dynamics reveal specific stage 2 vulnerability in insomnia. *Sleep*, zsx117.

Werth, E., Dijk, D. J., Achermann, P., & Borbély, a a. (1996). Dynamics of the sleep EEG after an early evening nap: experimental data and simulations. *The American Journal of Physiology*, *271*(3 Pt 2), R501–R510.

Wilson, S., & Argyropoulos, S. (2012). Sleep in schizophrenia: Time for closer attention. *British Journal of Psychiatry*, *200*(4), 273–274. https://doi.org/10.1192/bjp.bp.111.104091

Yetton, B. D., McDevitt, E. A., Cellini, N., Shelton, C., & Mednick, S. C. (2018). Quantifying sleep architecture dynamics and individual differences using big data and Bayesian networks. *PLoS ONE*, *13*(4). https://doi.org/10.1371/journal.pone.0194604

Yetton, B. D., Niknazar, M., Duggan, K. A., McDevitt, E. A., Whitehurst, L. N., Sattari, N., & Mednick, S. C. (2016). Automatic detection of rapid eye movements (REMs): A machine learning approach. *Journal of Neuroscience Methods*, *259*. https://doi.org/10.1016/j.jneumeth.2015.11.015

Yetton, B. D., Lacourse, K., Mednick, S. C., & Warby, S. C. (n.d.). Massive Online Data Annotation (MODA): crowdsourcing to generate high quality sleep spindle annotations from EEG data. *Nature Scientific Data*.

Young, T., Rabago, D., Zgierska, A., Austin, D., & Finn, L. (2003). Objective and subjective sleep quality in premenopausal, perimenopausal, and postmenopausal women in the Wisconsin Sleep Cohort Study. *Sleep*, *26*(6), 667–672.

Zhang, L., Samet, J., Caffo, B., & Punjabi, N. M. (2006). Cigarette smoking and nocturnal sleep architecture. *American Journal of Epidemiology*, *164*(6), 529–537. https://doi.org/10.1093/aje/kwj231